Verb system and verb usage in spoken and written Italian

Rosa Giordano[1], Miriam Voghera[2]

[1]Università di Perugia – Italia; CIRASS - Università Federico II di Napoli – Italia

[2]Università di Salerno – Italia

Abstract

It has been noted that spoken and written registers differ as far as the usage of verbs is concerned. In this article we investigate these differences between the two modalities of discourse taking into account lexical and grammatical factors.

First of all, we analyse in detail the frequency of verb types, verb forms and verb tokens in spoken and written texts. Secondly we investigate the frequency of usage of different verbal categories: mood, tense, person, number. Since Italian verb has a very high number of inflected forms (about fifty), we explore if, and to what extent, the difference in the use of verbs in spoken and written texts can be attributed to a different internal organisation of verbal system in the two modalities.

We use data coming from the two greatest available Italian corpora: the *Lessico di frequenza dell'italiano parlato* (LIP), for spoken Italian, and the *Lessico di frequenza della lingua italiana contemporanea* (LIF), for written Italian.

The data will show that spoken and written texts present more differences at the lexical level than at the grammatical level.

Keywords: spoken language, written language, lexicon, grammar, verbal system, Italian, frequency, verb types, verb forms, verb tokens.

1. Introduction

Although we have several studies on the differences of the use of parts of speech in spoken and written language (Halliday, 1985; Biber, 1995; Miller and Weinert, 1998; LGSWE, 1999; Voghera, in print a, b), we still have few data on the actual usage of the verb in texts of different registers. In this article we present a quantitative analysis of the Italian verb system in spoken and written texts, taking into account both lexical and grammatical factors.

First of all, we investigated if the different use of verbs in spoken and written texts could be attributed to lexical organisation, that is to differences in the frequency of verb types, verb forms and verb tokens in the two modalities. Secondly, we considered the relationship between verb type and verb forms and the frequency of different verb forms and tokens to understand if the verbs are differently used in spoken and written texts with respect to the grammatical meaning they convey. Therefore, we tagged all the verbs by mood, tense, person and number to design the distribution and the frequency of verbal categories in spoken and written language.

2. Methodology and data

A preliminary terminological explanation is needed. The following terms will be used: 1) *verb type* to refer to the lexical item to which every verbal inflectional form is assigned; 2) *verb*

form to refer to the different inflected forms which a verb type can assume; 3) *verb token* to refer to the occurrence of the different forms of a verb type.

We have collected data from two frequency dictionaries of Italian. The spoken data come from *Lessico di frequenza dell'italiano parlato* (LIP) (De Mauro et al., 1993) which is based on a corpus consisting of about 500.000 word tokens representing five different types of texts: a) face-to-face conversations; b) telephone conversations; c) non-free dialogical interactions; d) monologues (lectures, sermons…); e) radio and TV programmes. The third group includes texts in which the dialogical interaction does not proceed without restraint, but is guided by one of the speakers: typical texts are interviews, classroom interactions and debates. Each register is represented by samples of texts amounting approximately to 100.000 word tokens; texts were recorded in four different cities to represent the diatopic variety of Italian language[1]. The written data come from the *Lessico di frequenza della lingua italiana contemporanea* (LIF) (Bortolini et al., 1971), a frequency dictionary of written Italian. The LIF is based on a corpus consisting of about 500.000 word tokens representing five different types of texts: a) newspapers and magazines; b) textbooks for primary school; c) novels; d) theatre plays; e) movie scripts.

Both LIP and LIF present frequency lists ranked by frequency of usage. The LIP lists report all the types, that is types with frequency ≥ 1 and usage ≥ 0. The LIF lists include only types with frequency ≥ 3 and usage ≥ 1.78. In order to make comparable the two lists, we have not considered the types that in the LIP list have frequency < 3 and usage < 2.

We extracted the frequency of verbs included in the first 2000 word types of both frequency dictionaries. This range of frequency represents the so-called basic vocabulary (*vocabolario fondamentale*), which is supposed to be the core of the lexical knowledge. Summing the two corpora, the word tokens corresponding to the first 2000 word types are 905.551 out of 1.000.000, and they cover 91.1% of the LIP corpus and 90.7% of the LIF corpus. From all these data we can assume that verb frequency in the basic spoken and written vocabulary is representative of the general trend in speaking and writing.

3. The lexicon

First of all we considered the lexical richness of LIP and LIF, i.e. the number and distribution of verb types. In table 1 we show the percentage of verb types out of the total number of word types and out of the total number of verb types in each corpus.

Verb Types	***LIP***			***LIF***		
	No	***% T***	***% V***	***No***	***% T***	***% V***
Rank Range 0-2000	425	21.2	53.4	559	27.9	43.3
Rank Range 2001-forward	371	14.4	46.6	732	23.7	56.7
TOTAL	796	17.2	100	1291	24.1	100

Table 1 – Number of verb types of the whole corpus for LIP and for LIF; the percentages are calculated out of all the word types (% T) and out of all the verb types (% V).

[1] Italian presents wide and deep regional differences, which however can be ignored for the aim of this paper.

The spoken and written data show several differences. The LIF presents a greater number of verb types: on average, every 3 verb types in the LIP we find 4 verb types in the LIF. Besides, there is a difference in the type distribution: the rank from 0 to 2000 contains 53% of the total verbal types of the LIP but just 43% of the LIF. This means that the *basic vocabulary* has a major relevance in spoken texts than in written ones. This is also confirmed by the fact that the verb types included in the *basic vocabulary* cover the 94% (= 80252 tokens) of the total verbal tokens in the LIP texts and the 92.5% (= 83855 tokens) of the total verbal tokens in the LIF texts.

Surprisingly, we found only 359 types in common to the two corpora, which cover 84.5% of the spoken verb types and only 64.2% of the written verb types. These figures confirm the major richness and the variety of the written vocabulary, which mainly depend on diachronic, diaphasic and diamesic factors. The LIF corpus includes texts published between 1945 and 1968, while the LIP corpus was collected between 1990 and 1993. Secondly, spoken and written registers tend to select different topics and different levels of formality. This explains why verbs such as *'patire'* ('to be affected by pain or grief') or *'recare'* ('to convey') are unlikely to appear in conversation, but are common in the written corpus. A semantic analysis of verb types included only in the LIF and of those included only in the LIP will allow to discern which factor is the most relevant.

4. The grammar

In order to evaluate if the usage of verbs in written and spoken texts depends on differences in the organisation of the verbal system, we report here two different set of data. In section 4.1. we examine the relationship between forms and types in LIP and LIF corpora; in section 4.2. we show the frequency of usage of the categories of the Italian verb: mood, tense, person, number.

4.1. Verb types and verb forms

The total number of verb forms is quite different in the two corpora; there are 9718 verb forms in the LIF corpus and 5862 in the LIP corpus: so each LIF verb type presents, on average, about 3 forms more than a LIP verb type. The forms/types ratio is, in fact, of 13.7 for the LIP and of 17.1 for the LIF.

LIP	***Conversations***	***Telephone***	***N-f dialogues***	***Monologues***	***Radio and TV***	***Average***
Tokens	17742	18548	14753	13703	15504	16050
Forms	2354	2184	2602	2604	2548	2465
Types	398	378	414	415	415	404
Ratio F/T	5.9	5.7	6.3	6.3	6.1	6.1

Table 2 – Number of verb tokens, verb forms and verb types per kind of text in the LIP corpus, and average values of forms/type ratio.

LIF	*Newspapers*	*Textbooks*	*Novels*	*Theatre plays*	*Movie scripts*	*Average*
Tokens	20474	15530	22963	12376	12512	16771
Forms	4087	4305	3751	3646	3492	3854
Types	548	552	535	545	531	542
Ratio F/T	7.4	7.8	7	6.7	6.6	7.1

Table 3 – Number of verb tokens, verb forms and verb types per kind of text in the LIF corpus, and average values of forms/type ratio.

In tables 2 and 3 we present the LIP and LIF data per kind of texts and their average values. Comparing these data to global ones, it is possible to notice two different trends for types and forms. The average of types is 404 in the LIP and 542 in the LIF, which corresponds nearly to 95% of verb types in both corpora. Since only 42% (2465) of forms are used on average in the spoken texts and 40% (3854) in the written texts, the ratio forms/type is 6.1 in the LIP and 7.1 in the LIF.

If we look at the forms/type ratio in relation to the rank range of the verb types, we see that the lower the rank the lesser the number of forms. Table 4 shows that both LIP and LIF lists present higher number of forms per type in the first rank range, while this number noticeably decreases in the second range. Therefore spoken and written language reveal a similar trend in the distribution of forms related to frequency rank, although the LIF presents always more forms.

We computed the difference between the number of forms per type in LIP and LIF using the lists of the verbs present in the two corpora (see 3). Only 4.5% of types has the same number of forms in LIP and LIF; 81.3% has more forms in the LIF and just 14.2% has more forms in the LIP[2].

All these data confirm the similar organisation of the structure of the verbal system in the two modalities, since the proportions of types and forms actually used are similar in the two corpora; on the contrary, the major quantity and variety of forms of the system clearly emerge in written language.

	Type Rank Range 1-1000	***Type Rank Range 1001-2000***
Forms/type ratio LIP	19.0	8.4
Forms/type ratio LIF	22.8	11.5

Table 4 – Forms/type ratio per rank range for LIP and LIF.

[2] It is interesting to note that the first 20 verbs ordered by number of forms and by rank are the same in both LIP and LIF lists: *fare* (to do), *dire* (to tell), *dare* (to give), *andare* (to go), *vedere* (to see), *avere* (to have), *mettere* (to put), *tenere* (to hold/to keep/to bear), *prendere* (to take), *essere* (to be), *lasciare* (to let), *potere* (can/to be able to), *dovere* (must/to have to), *stare* (to stay), *trovare* (to find), *volere* (to want), *portare* (to bring), *sapere* (to know), *parlare* (to talk/to speak), *sentire* (to feel/to hear), *venire* (to come). This correspondence disappears as soon as we go to lower ranks.

4.2. The categories of verb

We have counted the forms of each verb type and have extracted their total frequency. Each form was tagged in order to classify it by mood, tense, person and number. We have considered all the simple forms of the verbs, excluding compound forms with auxiliaries, since LIP and LIF count separately the frequency of auxiliaries and the frequency of past participles.

4.2.1. Mood

Data on mood show two major characteristics (table 5). Moods have similar distribution both in LIP and LIF: indicative and infinitive are the most frequent, conditional, subjunctive, imperative and gerund the least frequent. Indicative presents the greatest number of forms and tokens in both corpora, but LIP shows about 6% of tokens of indicative more than LIF. These data confirm previous studies on the distribution of verb moods of Italian, according to which in spoken texts indicative tends to replace subjunctive in contexts in which the latter could be requested (Voghera, 1992).

Forms and tokens		***Indicative***		***Subjunctive***		***Conditional***		***Imperative***		***Infinitive***		***Gerund***	
		F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
LIP	***No***	1924	42529	201	1463	171	1454	272	780	1096	12233	307	1223
	%	32.8	53	3.4	1.8	2.9	1.8	4.6	1	18.7	15.2	5.2	1.5
LIF	***No***	3338	38971	445	1592	231	1312	423	1352	1965	13856	537	1347
	%	34.4	46.5	4.5	1.9	2.4	1.6	4.4	1.6	20.2	16.5	5.5	1.6

Table 5 – Number and percentage of forms (F) and tokens (T) per mood in the LIP and in the LIF. Participles and homograph forms are excluded.

LIP	***Conversations***		***Telephone***		***N-f dialogues***		***Monologues***		***Radio and TV***		***TOTAL***	
	F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
Indicative	31.8%	53.8%	36.3%	58.1%	32.7%	50.0%	34.1%	49.1%	32.7%	52.2%	32.8%	53.0%
Subjunctive	3.3%	1.9%	3.0%	1.5%	2.7%	2.0%	2.1%	1.8%	2.7%	1.9%	3.4%	1.8%
Conditional	2.9%	1.9%	2.9%	1.6%	2.8%	2.3%	2.8%	1.6%	2.9%	1.7%	2.9%	1.8%
Imperative	4.0%	1.0%	4.9%	1.6%	2.1%	0.6%	2.1%	0.5%	3.1%	1.0%	4.6%	1.0%
Infinitive	20.3%	14.7%	19.3%	11.2%	19.4%	18.0%	18.9%	18.5%	20.7%	15.2%	18.7%	15.2%
Gerund	3.7%	1.4%	3.5%	1.0%	5.7%	2.1%	4.3%	1.7%	4.3%	1.6%	5.2%	1.5%

Table 6 – Percentages of forms and tokens of moods per text in the LIP. Participles and homograph forms are excluded.

LIF	*Newspapers*		*Textbooks*		*Novels*		*Theatre plays*		*Movie scripts*		*TOTAL*	
	F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
Indicative	29.4%	45.4%	35.7%	49.1%	28.8%	45.3%	30.8%	39.8%	40.7%	53.9%	34.4%	46.5%
Subjunctive	3.2%	1.8%	4.2%	2.8%	3.8%	1.6%	3.2%	2.2%	3.0%	1.4%	4.5%	1.9%
Conditional	3.2%	2.2%	1.4%	1.0%	2.8%	1.9%	1.8%	1.7%	0.8%	0.5%	2.4%	1.6%
Imperative	5.1%	2.5%	2.0%	0.8%	5.5%	2.6%	0.8%	0.3%	1.5%	0.6%	4.4%	1.6%
Infinitive	23.0%	17.0%	21.1%	18.1%	23.3%	16.1%	19.8%	18.8%	16.1%	12.4%	20.2%	16.5%
Gerund	3.1%	1.0%	5.7%	2.6%	2.1%	0.6%	5.0%	2.3%	5.2%	2.6%	5.5%	1.6%

Table 7 – Percentages of forms and tokens of mood per text in the LIF. Participles and homograph forms are excluded.

Tables 6 and 7 report data on the different kinds of texts of each corpus. We limit our considerations to some points relevant to the aim of this paper. Using the terms proposed by Nencioni (Nencioni, 1983) in a classical essay on the difference between spoken and written Italian, we see that indicative has the highest percentages in the *parlato-parlato*, i.e. conversations, and, among the written registers, in the *scritto-parlato,* i.e. movie scripts, which are texts written to reproduce spoken conversation. It is interesting to point out that the frequency of indicative decreases with the degree of formality of texts, while the frequencies of infinitive, gerund and subjunctive seem to be higher in formal texts. This is probably related to a more frequent use of subordination in this kind of texts, since infinitive, gerund and subjunctive are often used in subordinate clauses.

4.2.2. Tense

Tense and aspect are not separated in Italian inflection, therefore what we tagged here as tense distinctions is actually an expression of both tense and aspect (Bertinetto, 1986).

In both LIP and LIF the present is the most used tense, although LIP has more present tokens than LIF, as we can see in tables 8 and 9. In the LIP nearly 70% of presents is indicative; in the LIF the present indicative is about 60%. The past tenses are more frequent in the written texts, but there is a great difference between imperfect and past simple (past historic). The imperfect has 9.6% of the LIF tokens and 6.3% of the LIP tokens; the simple past has 4.9% of LIF tokens, but only 0.4% of the LIP tokens. The future has basically the same frequency in both LIP and LIF.

We can note several differences among the different kinds of text. The present indicative is considerably less frequent in textbooks (17%), while in the same texts the simple past reaches nearly 10% of tokens. Monologues and radio and TV programs present the highest percentage of tokens of future (3%), while conversations present the lowest (1.5%). It is interesting to note that in the LIF we find the highest percentage of simple past (15.1%) in the movie scripts, which are supposed to reproduce conversation, but are far from presenting the same distribution of tense, as it is also shown by the percentage of present tokens, which is twenty per cent lower (45.3%).

LIP		Conversations		Telephone		N-f dialogues		Monologues		Radio and TV		TOTAL	
		F	T	F	T	F	T	F	T	F	T	F	T
Present	Ind	18.9%	46.0%	20.2%	50.1%	17.6%	39.8%	18.0%	40.5%	18.0%	44.8%	14.6%	44.6%
	Subj	1.2%	1.5%	1.0%	1.0%	1.2%	1.5%	1.2%	1.3%	1.3%	1.4%	1.1%	1.3%
	Cond	3.2%	1.9%	3.0%	1.6%	2.6%	2.3%	2.0%	1.6%	2.7%	1.7%	2.9%	1.8%
	Imp	4.0%	1.0%	4.9%	1.6%	2.1%	0.6%	2.0%	0.5%	3.0%	1.0%	4.7%	1.0%
	Inf	20.3%	14.7%	19.2%	11.2%	19.3%	18.0%	18.8%	18.5%	20.6%	15.2%	18.7%	15.2%
	Ger	3.7%	1.4%	3.5%	1.0%	5.7%	2.1%	4.2%	1.7%	4.2%	1.6%	5.2%	1.5%
	TOT	65.8%	66.4%	65.8%	66.6%	62.9%	64.2%	62.3%	64.0%	65.0%	65.8%	47.6%	65.5%
Imperfect	Ind	7.4%	6.1%	8.6%	6.1%	8.8%	7.9%	7.3%	4.9%	6.5%	4.1%	8.4%	5.8%
	Subj	1.7%	0.5%	1.9%	0.4%	1.7%	0.6%	1.5%	0.5%	1.6%	0.5%	2.3%	0.5%
	TOT	9.3%	6.6%	10.6%	6.5%	10.6%	8.5%	9.0%	5.4%	8.3%	4.6%	10.7%	6.3%
Future		4.6%	1.5%	5.8%	1.6%	4.6%	1.7%	6.7%	3.1%	7.4%	3.0%	7.3%	2.1%
Simple past		0.8%	0.2%	1.6%	0.4%	2.3%	0.6%	2.6%	0.6%	1.4%	0.2%	2.1%	0.4%

Table 8 – Percentages of forms and tokens of tenses per text in the LIP. Participles and homograph forms are excluded.

LIF		Newspapers		Textbooks		Novels		Theatre plays		Movie scripts		TOTAL	
		F	T	F	T	F	T	F	T	F	T	F	T
Present	Ind	14.1%	36.3%	9.9%	17.8%	13.5%	37.3%	13.6%	27.6%	12.9%	28.9%	10.7%	30.8%
	Subj	1.0%	0.7%	0.5%	0.5%	1.2%	0.7%	1.2%	1.3%	0.5%	0.3%	1.2%	0.7%
	Cond	3.2%	2.2%	1.4%	1.0%	2.8%	1.9%	1.8%	1.7%	0.8%	0.5%	2.4%	1.6%
	Imp	5.1%	2.5%	2.0%	0.8%	5.5%	2.6%	0.8%	0.3%	1.5%	0.6%	4.4%	1.6%
	Inf	23.0%	17.0%	21.1%	18.1%	2.3%	16.1%	19.8%	18.8%	16.1%	12.4%	20.6%	16.5%
	Ger	3.1%	1.0%	5.7%	2.6%	2%	0.6%	5%	2.3%	5%	2.6%	5.5%	1.6%
	TOT	63.1%	59.6%	45.4%	40.8%	63.4%	59.2%	50.8%	52.0%	44.7%	45.3%	44.1%	52.8%
Imperfect	Ind	6.8%	5.1%	14.3%	20.2%	7.2%	4.7%	7.1%	5.9%	10.1%	8.5%	10%	8.4%
	Subj	2.2%	1.0%	3.7%	2.2%	2.6%	0.9%	2.0%	0.9%	2.5%	1.1%	3.4%	1.2%
	TOT	9.0%	6.1%	18.0%	22.3%	9.8%	5.6%	9.1%	6.8%	12.5%	8.2%	13.4%	9.6%
Future		6.3%	3.2%	2.6%	1.4%	6.4%	2.8%	3.9%	2.7%	3.5%	1.4%	5.5%	2.4%
Simple past		2.7%	0.7%	9.5%	9.6%	1.8%	0.4%	6.2%	3.7%	14.5%	15.1%	8.1%	4.9%

Table 9 – Percentages of forms and tokens of tenses per text in the LIF. Participles and homograph forms are excluded.

4.2.3. Person and Number

Tables 10 and 11 show the frequency and the distribution of the categories of person and number in the verb forms and tokens.

Neither the verb forms nor the verb tokens show a different distribution in written and spoken usage, as far as the category of person and number are concerned. The third person is the most frequent, while the second person is the least frequent; singular is most frequent than plural in both written and spoken texts, and in every text.

Forms and tokens		***1***		***2***		***3***	
		F	***T***	***F***	***T***	***F***	***T***
LIP	***No***	952	13555	826	7168	1592	31332
	%	16.2	16.9	14.1	8.9	27.1	39
LIF	***No***	1251	11885	1254	8749	2968	32727
	%	12.9	14.2	12.9	10.4	30.5	39

Table 10 – Distribution of the category of person in LIP and LIF verb forms and tokens. Homograph forms are excluded.

Forms and tokens		***Singular***		***Plural***	
		F	***T***	***F***	***T***
LIP	***No***	1948	43510	1526	11999
	%	33.2	54.2	26	15
LIF	***No***	3533	44827	2146	9868
	%	36.3	53.5	22.1	11.8

Table 11 – Distribution of the category of number in LIP and LIF verb tokens. Homograph forms are excluded.

Data reported in tables 12, 13, 14 and 15 on the frequency of person and number categories in each kind of spoken and written text allow a detailed analysis. The higher frequency of the third person in both spoken and written texts can be explained by the higher frequency of copula *è* (is), the presentative construction *c'è* (there is), and by the use of impersonal forms. The frequency of the first person is naturally higher in spoken texts (16.9%), and among them in telephone conversations. However, we found some unexpected data, such as the high frequency of first person in newspapers, and the higher frequency of the second person in written texts.

The contexts must be analyzed in detail, at the moment we can only suppose that both these data can be explained by the presence of dialogical parts.

LIP		*Conversations*		*Telephone*		*N-f dialogues*		*Monologues*		*Radio and TV*		*TOTAL*	
		F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
1	***Sg***	10.2%	11.2%	12.5%	15.0%	6.3%	7.9%	6.7%	6.0%	8.8%	10.9%	8.3%	10.5%
	Pl	6.5%	4.5%	7.2%	4.2%	7.2%	7.1%	7.6%	8.7%	8.4%	8.3%	7.9%	6.3%
	TOT	16.7%	15.7%	19.7%	19.2%	13.5%	15.0%	14.3%	14.7%	17.2%	19.2%	16.2%	16.9%
2	***Sg***	9.4%	7.4%	10.7%	11.3%	4.3%	3.2%	4.1%	1.9%	6.4%	5.4%	7.7%	6.2%
	Pl	4.5%	1.4%	3.8%	1.1%	4.5%	2.0%	5.7%	4.5%	6.9%	5.3%	6.4%	2.7%
	TOT	13.9%	8.8%	14.5%	12.4%	8.9%	5.2%	9.8%	6.4%	13.3%	10.7%	14.1%	8.9%
3	***Sg***	16.3%	35.1%	15.7%	33.6%	19.0%	32.7%	18.9%	32.7%	15.6%	31.3%	15.5%	33.2%
	Pl	9.3%	4.9%	9.8%	4.3%	11.9%	8.1%	12.3%	7.6%	9.2%	5.3%	11.6%	5.9%
	TOT	25.6%	40.0%	25.5%	37.9%	30.9%	40.8%	31.3%	40.3%	24.8%	36.6%	27.1%	39.1%

Table 12 – Percentages of forms (F) and tokens (T) of person per text in the LIP. Participles and homograph forms are excluded.

LIF		*Newspapers*		*Textbooks*		*Novels*		*Theatre plays*		*Movie scripts*		*TOTAL*	
		F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
1	***Sg***	11.2%	16.3%	9.3%	9.9%	9.7%	14.3%	5.6%	6.5%	2.2%	5.2%	8.5%	11.5%
	Pl	4.2%	2.9%	3.0%	1.6%	5.2%	3.8%	3.3%	2.3%	2.8%	2.4%	4.4%	2.7%
	TOT	15.4%	19.2%	12.3%	11.5%	14.9%	18.1%	8.9%	8.8%	5.0%	7.6%	12.9%	14.2%
2	***Sg***	11.1%	11.6%	5.0%	4.3%	11.5%	14.3%	2.0%	0.8%	4.1%	4.5%	7.9%	8.3%
	Pl	5.9%	3.1%	2.1%	0.8%	5.4%	2.7%	3.2%	1.4%	3.2%	4.0%	5%	2.1%
	TOT	17.1%	14.7%	7.1%	5.1%	16.8%	17.0%	5.1%	2.3%	7.3%	8.5%	12.9%	10.4%
3	***Sg***	14.9%	28.2%	23.3%	37.4%	15.1%	27.9%	21.6%	33.2%	25.1%	38.2%	17.9%	32.1%
	Pl	7.5%	3.7%	9.1%	6.1%	8.4%	3.7%	11.9%	9.1%	20.0%	17.3%	12.6%	7.0%
	TOT	22.4%	31.8%	32.5%	43.5%	23.5%	31.7%	33.5%	42.4%	45.1%	55.5%	30.5%	39.0%

Table 13 – Percentages of forms and tokens of person per text in the LIF. Participles and homograph forms are excluded.

LIP	*Conversations*		*Telephone*		*N-f dialogues*		*Monologues*		*Radio and TV*		*TOTAL*	
	F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
Singular	37.8%	57.8%	40.5%	63.2%	31.3%	48.7%	31.2%	45.5%	32.4%	52.3%	33.2%	54.2%
Plural	20.4%	10.8%	20.8%	9.6%	23.6%	17.2%	25.6%	20.8%	24.5%	18.8%	26%	15.0%

Table 14 – Percentages of forms and tokens of number per text in the LIP. Participles and homograph forms are excluded.

LIF	*Newspapers*		*Textbooks*		*Novels*		*Theatre plays*		*Movie scripts*		*TOTAL*	
	F	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***	***F***	***T***
Singular	52.8%	58.2%	54.8%	52.9%	52.4%	58.7%	49.8%	42.2%	46.3%	48.3%	36.3%	53.5%
Plural	21.0%	9.6%	18.4%	8.5%	22.2%	10.2%	25.4%	12.9%	32.4%	21.3%	22.1%	11.8%

Table 15 – Percentages of forms and tokens of number per text in the LIF. Participles and homograph forms are excluded.

6. Conclusions

Both lexical and grammatical data allow several reflections on verb system and usage in spoken and written Italian.

First of all we noticed that spoken texts show a higher degree of similarity in the use of verbs, while there is a greater variety of use among the written texts. This probably depends on the context of production of written texts, which allows a process of text editing tending to more accurate and varied linguistic choices. Moreover, there are external factors, such as the ancient tradition of *genre* distinction of written texts, which determine a more varied usage both in lexicon and grammar. Differently, spoken texts show a smaller amount of variation and tend to reproduce the same linguistic choices. Since spoken texts are not planned or previously organised, the speaker must organise and utter the text simultaneously. This on-line production constraints do not allow to choice among alternatives, and produce a lot of repetition both in lexicon and grammar. This is the reason that determines a stronger similarities among spontaneous spoken texts compared with the written ones.

The analysis of tokens, types and forms number seems to confirm that written texts show a greater lexical variety. The written corpus presents a higher number of types and forms, and higher forms/type ratio, and this trend is directly proportional to the degree of formality of texts. Moreover, the verbs included in the *basic vocabulary* cover only 64.2% of the written verb types, but 84.5% of the spoken verb types. This means that when we speak we tend to use very often a smaller group of verb types.

Spoken and written corpora present strong similarities in the grammar of the verb usage. Although spoken texts tend to reduce the number of verb categories, the system of moods does not present any difference: the total percentages of usage of different moods are basically the same. The only difference we found is the higher number of different forms per mood in written corpora.

The tense system presents more differences between the spoken and written corpora, even because in Italian inflection tense distinctions are related to aspect distinctions. The simple past is basically absent in spoken texts and even the imperfect is less used than in written texts.

The data on person and number are not easily comprehensible. The third singular verb forms are the most frequent both in spoken and written texts. The first singular is more frequent in spoken texts, but is more frequent in newspapers than in theatre plays; the second singular is more frequent in written texts, but not, as expected, in movie scripts and theatre plays.

To summarize, most of these data confirm a general greater lexical richness of written texts and a tendency of spoken texts to prefer unmarked and basic linguistic structure. In fact, we

found that a relatively small amount of verb types and verb forms occurs very often in spoken texts, while written texts tend to use less frequently a greater amount of verb types and verb forms. It is interesting to remind that typical spoken texts show a general higher frequency of verb tokens than typical written texts. Therefore spoken texts use the verbs much more often than written texts, although the lexicon and the grammar of verbs are in some way reduced in spoken language. This apparent paradox can be explained by the fact that the higher frequency of verb tokens in spoken texts does not depend on lexical or grammatical factors.

According to Biber (Biber, 1995) and Voghera (Voghera, in print b) the higher number of verb tokens in spoken registers must be attributed to the typical syntactical and textual strategies used in speaking. The need to recall portions of text without the support of an external memory determines a typical feature of spoken utterances: they tend to reproduce the sequence of events structuring information in serial patterns. Thus, the quantity of information develops through an additive process, which can easily be reconstructed, even in case of project changes or interruptions. Syntactically, this means short clauses, and this often implies absence of nominal constituents or very simple nominal constituents. It has been noted that in spoken sentences verb valences can be saturated by pronouns or simple NPs, because the semantic and syntactic relations can easily be reconstructed by making appeal to contextual cues (Miller and Weinert, 1998). This is one of the most important factors to determine a higher frequency of verbs.

References

Bertinetto P. M. (1986). *Tempo, aspetto e azione nel verbo italiano*. Firenze, Accademia della Crusca.

Biber D. (1995). *Dimensions in Register Variation. A Cross-Linguistic Comparison*. Cambridge, Cambridge University Press.

Halliday M.A.K. (1985). *Spoken and Written Language*. Oxford, Oxford University Press.

LGSWE: Biber D., Johansson S., Leech G., Conrad S. and Finegan E. (1999). *Longman Grammar of Spoken and Written English,* London, Longman.

LIF: Bortolini U., Tagliavini C. and Zampolli A. (1972). *Lessico di frequenza della lingua italiana contemporanea*. Milano, Garzanti-IBM.

LIP: De Mauro T., Mancini F., Vedovelli M. and Voghera M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano, Etaslibri.

Miller J. and Weinert R. (1998). *Spontaneous Spoken Language. Syntax and Discourse*. Oxford, Clarendon Press.

Nencioni G. (1983). "Parlato-parlato, parlato-scritto, parlato-recitato". In Nencioni G., *Di scritto e di parlato*. Bologna, Zanichelli.

Voghera M. (1992). *Sintassi e intonazione nell'italiano parlato*. Bologna, Il Mulino.

Voghera M. (in print a). La distribuzione delle parti del discorso nel parlato e nello scritto. In van Deyk R., editor, *La variabilité en langue*, Tübingen, Gunter Narr Verlag.

Voghera M. (in print b). Nouns and verbs in speaking and in writing. In Burr E., editor, *Proc. of SILFI 2000 (Società Internazionale di Linguistica e Filologia Italiane).*