The LIPS Corpus (Lexicon of Spoken Italian by Foreigners) and the acquisition of vocabulary by learners of Italian as L2

Francesca Gallina
University for Foreigners of Siena, Italy

Abstract

The aim of this paper is to present corpus-based research on the acquisition of the vocabulary of Italian as L2. The goal of the research was to study the lexical uses of non-native speakers and the processes of lexical acquisition underlying these uses. The informants of the corpus were non-native speakers learning Italian both within Italy and outside of it in order to compare the development of lexical competence in different learning contexts. The main results show how lexical competence develops above all quantitatively at the beginning and intermediate levels, as well as how it develops qualitatively at the more advanced levels in particular. Different learning inputs greatly affect the development of lexical competence: learners acquiring Italian in Italy have a deeper knowledge of the Italian lexicon compared to learners learning Italian outside of Italy.

Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching, Vol. 4: Papers from LAEL PG 2009
Edited by Grace Bota, Helen Hargreaves, Lai Chia-Chun & Rong Rong

Introduction

I present here corpus-based research on the acquisition of the vocabulary of Italian as a second language carried out at the Centro di Eccellenza della ricerca – *Osservatorio Linguistico permanente dell'italiano diffuso fra stranieri e delle lingue immigrate in Italia* of the University for Foreigners of Siena.

This research has two main goals: on the one hand, to study the lexical uses of learners of Italian as a second language and the lexical choices they make; on the other, to study the processes of lexical acquisition and development underlying these uses.

The learning of vocabulary depends on a lot of factors – linguistic, extralinguistic and individual. It is a creative, incremental, dynamic and continuous process. So, in order to study the acquisition of vocabulary we need to keep in mind many aspects: implicit and explicit vocabulary learning (Ellis N. 1994; Ellis R. 1995, 1995), the structure of the mental lexicon (Schreuder & Weltens 1993; Singleton 1999; Wolter 2001), the difference between active and passive vocabulary (Laufer 1998; Laufer & Paribakht 1998; Meara 1990, Melka 1997; Mondria & Wiersma 2004; Nation 2001; Palmberg 1989), the multidimensional model of lexical competence development (Haastrupt & Henriksen 2000; Henriksen 1999; Meara 1996, 1999; Read 2004; Schmitt & McCarthy 1997).

I assumed that learning input is a crucial factor in determining which words and how many enter into the second language vocabulary, especially at the beginner levels. I therefore compared the lexical competence of learners exposed to different learning inputs. I also assumed, following Meara (1996), that the size dimension of lexical competence grows faster mainly at the beginning and intermediate levels. I thus compared the lexical competence of learners at different levels. Observing the lexical uses and their underlying acquisition/learning processes, I tried to hypothesize some general development lines for lexical competence and to study the role of input in different learning contexts.

The LIPS (Lexicon of Spoken Italian by Foreigners) corpus

The research is based on a learner corpus of spoken data, the LIPS (Lexicon of Spoken Italian by Foreigners), drawn from the proficiency tests of the Certification of Italian as a Foreign Language (CILS) of the University for Foreigners of Siena, which is one of the proficiency centres for Italian as a foreign language[1]. Ball (2001, 2002), Barker (2004), and Read (2005) created corpora drawn from the proficiency tests of EFL using IELTS and ESOL tests to collect the corpora for their research.

The LIPS corpus consists of 1500 exams of around 500 candidates who took the tests between 1993 and 2006. The candidates of the corpus are non-native speakers learning Italian both within Italy and outside of it. The corpus consists of around

[1] On the objectives, contents and results of the LIPS corpus see also Bagna, *et al*. (2003), Carloni & Vedovelli (2005), Vedovelli (2006), Barni & Gallina (2008, 2009). On the CILS certification see Vedovelli (2005).

700,000 tokens and around 100 hours of spoken data. It is at present one of the biggest corpora of Italian as spoken by non-native speakers. The large dimension of the corpus makes it more reliable, as I can choose only those exams that are more similar to the spontaneous spoken language for mine research. For this research, I selected only 127 exams from 32 examinees from the LIPS: 40 exams from 10 examinees who took all the exams in Italy; 72 exams from 19 examinees who took all the exams outside of Italy; 11 exams from 3 examinees who took the exams both in and outside of Italy. This corpus consists of 62 499 tokens, 7075 types and 4646 lemmas. The results I had from the analysis of this sub-corpus are very similar to the results I obtained on the whole LIPS corpus and so they can indicate some general tendencies of development of lexical competence.

The candidates of the corpus have been exposed to very different types of learning inputs and different learning and acquisition processes. I selected only those candidates that took at least 3 proficiency tests of different levels (from A1 to C2 in CEF (2001) terms), in order to study lexical progression along the acquisition continuum at different stages. I thus include in the corpus examinees with 3 or 4 exams. I decided to pick out only successful exams to represent the proficiency level in order to be sure that the examinees' levels are comparable. I selected tests taken in different sessions in order to have a wide range of texts, so that in the corpus I could refer to different domains, different contexts and different communicative situations even if they are all examination contexts. In the corpus there are a lot of different topics (around 300) not only at different levels, but also in different examination sessions. The topics are more linked to daily life at the beginner levels (A1-B2), and are more formal and professional at the advanced levels (C1-C2). I therefore have a wide range of lexical fields and this makes the corpus very rich and, above all, more reliable.

The corpus is very rich also because the texts present different speech genres. There are two oral texts in every CILS proficiency exam, one dialogue with the examiner and one monologue. This permits me to evaluate the learner's ability to move in the linguistic space of spoken Italian. We need to keep in mind we are dealing with an exam context, which means the speakers do not have the same role, since one is the examinee and the other is the examiner. However, in order to make the conversation as spontaneous as possible, the examiner is given the task of simulating natural communicative contexts. What I found throughout the analysis of the corpus is that the communication between the examiner and the examinee was not always easy to define in terms of the dichotomy dialogue vs monologue. Sometimes one text was not fully a dialogue or a monologue, and there are a lot of mixed texts which stand on a continuum, the limits of which are dialogue on one side and monologue on the other. In this way, texts even in an examination context are quite similar to spontaneous interactions in natural contexts.

The spoken data were recorded on a CD-ROM or an audiotape, transcribed and annotated with part-of-speech and lemma information using open source software. Every exam of the corpus is then recorded in a database with some additional information in order to select the exams for different kinds of research: the examination date (it permits an examinee's interlanguage competence development to be followed); the examination centre (it gives some information on the L1 of the examinee, and so

gives the possibility of carrying out contrastive analysis between examinees of different mother tongues); the proficiency level (it describes the proficiency level and is on a six-level scale from A1 to C2); the total number of exams of the same examinee (it allows me to compare examinees that have the same number of exams so as to explore individual differences in the learning processes); the speech genre (it specifies whether the exams consist of a dialogue, a monologue or a mixed text); the examination topic (it gives some general information on the topics of the two texts of every exam).

Transcription and lemmatization of the corpus

In the first phase of the research I transcribed the whole corpus. As the main aim of the research was a lexical analysis, I chose an orthographical transcription, without any other annotation, because it would have been difficult later on to use the software for lexical research if the transcription had been more complex. Both the production of the examiner and of the examinee was transcribed to permit a reconstruction of the interactive dynamics. I decided to respect as far as possible the original spoken text during the transcription, hence I did not standardize the examinee's speech even when there were a lot of phonetic, syntactic, morphologic and lexical phenomena which were not standard Italian. I encountered several difficulties in transcribing interlanguage data on the one hand and L1 and other L2 words on the other, so I was sometimes unable to transcribe sequences of more than one word.

In the second phase of the research I carried out the lemmatization of the corpus. I lemmatized the examinee's turns only. First of all I used an open source software, Schmid's TreeTagger (1994, 1995), a probabilistic part-of-speech tagger, which proved to have various advantages in the experimental phase of the corpus. It produces fewer errors with non-native speakers' production than other software I used in the experimental phase, even though the TreeTagger's training corpus is not a learner corpus. It gives back the lemmatized text in columns: the original text is in the first column, the part of speech in the second and the lemma in the third column (figure 1).

sono nata	VER:ppros	nascere
eh	INT	eh
nel	PRE:det	nel
in	PRE	in
Millenovecinquantasei	NOM	Millenovecinquantasei
vivo	VER:pres	vivo
a	PRE	a
nella	PRE:det	nel
periferia	NOM	periferia
di	PRE	di
Parigi	NGR	Parigi
sono	VER:pres	essere
professoressa	NOM	professore
di	PRE	di
francese	NOM	francese
in	PRE	in
un	DET:indef	un
liceo	NOM	liceo
a	PRE	a
Parigi	NGR	Parigi

Figure 1 Example of lemmatization

The vertical distribution of the text has the advantage of letting you always have the entire text in front of you, so it is possible to analyse a token and its lemma in their real context. Finally, TreeTagger is more suitable for the successive analysis – e.g. extracting a frequency list – that I intended to do. After the automatic lemmatization I did a manual correction on the whole corpus, because of the many errors made by TreeTagger. This correction was absolutely necessary to disambiguate the usual errors a tagger makes during the lemmatization of a text, of which there are even more when the text is produced by a non-native speaker.

The main problems of the automatic lemmatization that absolutely required manual correction concern:

- the disambiguation of homographs that can have the same part of speech or the same lemma (e.g. *parti = VERB partire/ NOUN parte*; *Argentina* = ADJ *argentino*);
- the tagging of foreign words and interlanguage expressions (e.g. *io ando* = *io vado*);
- the attribution of the grammatical categories to words such as pronouns, conjunctions and adverbs;
- the tagging of false starts, broken words, words with orthographical errors that were transcribed exactly as they were in the original speech.

A lot of difficulties derive from the fact that the software could not interpret a text that is syntactically or morphologically anomalous for a program that has a training corpus of native-speakers' production, but that can be understood by a native speaker.

Furthermore, I have a corpus of oral text, and taggers are not usually trained on oral texts, but rather on written texts, so it is very hard for software to tag speech production correctly because of the characteristics peculiar to oral texts. Besides the tagset of TreeTagger, I also used some other tags to annotate the corpus, as I wanted to point out some typical characteristics of mine corpus. As the use of multi-word lexical units is considered very native-like for non-native speakers, I used a special tag for them, because I were interested in analysing the ability to use this kind of word by non-native speakers. I annotated the foreign words from L1 and other L2 with a special tag (ESO) to highlight the presence of words that do not belong to Italian. I also used a special tag for interlanguage data (INTL) to mark the creative production of non-native speakers, whose interlanguage competence is not standard Italian, but something between their L1 and the target language in the learning process.

Analysis of lexical richness

After lemmatization I extracted frequency and usage lists that can be compared to the frequency lists of Italian used by native speakers and especially with the *Italian Frequent Lexicon* (LIF) (Bertolini *et al*., 1971), and the *Italian Spoken Lexicon* (LIP) (De Mauro *et al*., 1993), the former being a frequency list of written Italian by native speakers and the latter being a frequency list of spoken Italian by native speakers.

I used some intrinsic and extrinsic measures of lexical richness (Meara & Bell, 2001) to study the lexical variety, the lexical sophistication, and the lexical density of the corpus, in order to compare (a) learners with different proficiency levels on the one hand and (b) learners with different learning inputs on the other.

I then compared the sub-corpus of the LIPS with the *Vocabolario di Base* (VDB) (De Mauro, 1980). The VDB represents the Italian core vocabulary, consisting of around 7000 words, and was generated from the most frequently used words of the LIF corpus. I used the VDB to see how many of the most frequently used words non-native speakers use when they speak Italian, as well as to have an idea of what kind of words they use compared to the most frequent lexical uses of native speakers (also with reference to the different proficiency levels).

Results

Quantitative analysis

First of all I measured how many words there are in the different levels of the corpus and how many tokens/types and lemmas. Figure 2 below shows how many tokens and speakers there are in each level of the corpus.

	A1	A2	B1	B2	C1	C2
Tokens	415	1486	12627	14703	14244	19024
Speakers	1	4	31	31	29	31

Figure 2 Tokens and speakers in each level

The average number of tokens that learners at different proficiency levels produce grows from level A1 to level C2, supporting the hypothesis that size dimension is one of the key dimensions of development of lexical competence, especially at the beginning levels, and even if this is also partly due to the exam format, observing the number of words in the same amount of time we can see that beginner learners produce fewer tokens than advanced learners (see figure 3 below).

	A1	A2	B1	B2	C1	C2
Tokens	415	371.5	407.32	474.32	491.27	604.96

Figure 3 Average number of tokens in different proficiency levels

At the beginning levels the acquisition of the lexicon is more quantitative (number of words learned) than qualitative (knowledge of different characteristics, uses and meanings of words), and after a certain level the process is not only a question of the quantity of words a speaker knows, but also of the quality and depth of lexical knowledge (that is, a better knowledge of the characteristics of a word, the ability to use it in a context and to relate it to other words), and of automaticity (the ability to use and access the vocabulary (Meara, 1996, Schmitt & Mc Carthy, 1997: 104)).

It may be observed in figure 4 below that the average number of tokens every non-native speaker produces speaking in Italian is higher in learners learning Italian in Italy or both within Italy and outside of it.

	Out of Italy	Italy	Mixed
Tokens	474.4	490.52	563.66

Figure 4 Tokens in learners with different learning contexts

Learners learning Italian only outside of Italy produce fewer tokens than the others. This means that learners who learn Italian in context or with frequent possibilities of contact with the language and native speakers are able to use more tokens than other learners, regardless of the nature of these words. These results support the hypothesis on the role of the learning input in the development of lexical competence and in determining the ability of producing words for learners learning an L2 in different learning contexts.

These kinds of analysis do not enable to understand the complexity of the process of lexical development, and it is therefore necessary to apply other kinds of analysis aimed at measuring the lexical richness, looking at the nature and not only at the quantity of words learners know and use.

The frequency list

From the whole corpus I extracted a frequency list and a usage list, which takes into consideration not only the frequency of words, but also how much words are scattered in a corpus. The most frequently used words are function words like conjunctions and prepositions, and then content words. They are all part of the Italian core vocabulary (VDB). In the first 100 most frequently used words of the usage list there are: 9 adjectives, 16 adverbs, 11 conjunctions, 3 articles, 2 adverbial phrases, 4 interjections, 16 nouns, 11 prepositions, 10 pronouns and 18 verbs. Foreign words, interferences and interlanguage expressions have a very low use in the frequency list, as do multi-word lexical units. There are also a lot of typical spoken phenomena among the most frequent words of the list, such as interjections, just as there are in the frequency lists of native speakers.

Lexical variety: Types/Tokens Ratio and Guiraud's Index

To assess the lexical variety I applied the Types/Tokens Ratio (TTR) and Guiraud's Index[2]. These kinds of measurement are very sensitive to the length of the texts they are applied to and are not easy to apply to non-native speakers' texts (Broeder *et al*., 1993, Granger & Wynne, 1999, Laufer & Nation, 1995, Vermeer, 2000, 2004). This notwithstanding, they are still the most used measures of lexical richness, and this is one of the reasons I decided to apply them to mine corpus, and moreover I can achieve a more general view of the development of lexical competence from these kinds of measurement, which provide quantitative data alongside qualitative data, such as those obtained from extrinsic measures of lexical richness that compare non-native speakers' vocabulary to the vocabulary used by native speakers.

Data on TTR and Guiraud's index (figures 5 and 6 below) confirm the difficulties of applying these measurements to long texts and to non-native speaker texts. They show that lexical variety is larger at the beginner levels, but this is due more to the different lengths of the texts than to a supposed broader variety of beginner learners' lexical competence. The growth of the lexical variety increases progressively but it is not constant; it is more intense especially at the beginning and intermediate levels and less intense at the advanced levels. This suggests that at the advanced levels measurements based only on the relationship between types and tokens are not suitable for describing the competence of advanced learners, but that Ialso need qualitative analysis.

Level	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
TTR	0.43	0.36	0.17	0.17	0.17	0.15

Figure 5 Type/Token Ratio in different proficiency levels

[2] TTR (V/N) and Guiraud's index (V/√N) are measures of lexical richness, in which 'types' (V) are the number of different words, and 'tokens' (N) the total number of words of a text. This kinds of measures are based on the assumption that more proficient learners have a larger vocabulary knowledge that allows them to avoid repetition by using a more varied vocabulary.

Level	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
Guiraud's index	8.88	13.95	20.18	20.63	20.69	21.46

Figure 6 Guiraud's index in different proficiency levels

The lexical variety (see figures 7 and 8 below) is larger in learners learning Italian in Italy than in learners learning Italian outside of Italy: contact with native speakers is a fundamental aspect of the development of lexical competence, confirming the relevance of input in the learning process.

	Outside of Italy	**Italy**	**Mixed**
TTR	0.12	0.16	0.19

Figure 7 Type/Token Ratio in learners with different learning contexts

	Outside of Italy	**Italy**	**Mixed**
Guiraud's index	23.31	23.0	18.18

Figure 8 Guiraud's index in learners with different learning contexts

Learners learning Italian in Italy show a bigger lexical variety, especially as far as the TTR ratio is concerned, than learners who are not directly exposed to the learning input of native speakers in the Italian context. The Guiraud's index results are slightly different, but in mine opinion this is more linked to the sensitivity of this kind of analysis to the length of texts and other characteristics of texts than to other factors.

Lexical density

To analyse the lexical density, namely a high percentage of lexical or content words as compared to grammatical or function words, I counted the content and the function words, and then analysed the distribution of the grammatical categories in every level and in every group according to the learning context.

In the corpus there is 60.2% of content words and 39.8% of function words. What I found looking at the different levels and at the different learning input is that, as with lexical variety, lexical density is larger in learners learning Italian in Italy (see figure 9 below), and it increases progressively especially at the beginner and intermediate levels and less at the more advanced levels (see figure 10 below).

	Outside of Italy	**Italy**	**Mixed**
Content words	59.58%	61.42%	58.50%
Function words	40.21%	38.19%	41.15%

Figure 9 Content/function words in learners with different learning contexts

	A1	**A2**	**B1**	**B2**	**C1**	**C2**
Content words	57.92%	58.86%	60.17%	61.85%	60.56%	58.34%
Function words	41.81%	40.98%	39.56%	37.97%	38.98%	41.48%

Figure 10 Content/function words in learners with different proficiency levels

Texts produced by learners learning Italian in Italy are lexically denser than those by learners not exposed to spontaneous learning inputs. Content words progressively increase from level A1 to B2, and then decrease in levels C1 and C2. The drop in content words at these levels may perhaps mean that beyond a certain level of proficiency learners can not only use content words they have been learning, but can also use function words in a profitable way to connect content words they already know. This supports the hypothesis on the relevance of the qualitative dimension of lexical competence at the more advanced levels, whereas the size dimension is not completely adequate to distinguish beginner and intermediate learners from proficient learners. To expand lexical competence it is not enough to increase the number of words a learner knows, but it is also necessary to be able to use them in their appropriate context, to know their various aspects, to connect them through function words enhancing the fluency of the text in a way that resembles native speakers' use, and to access them quickly. This is a further confirmation that the size dimension of lexical competence development is more important at the beginner levels, while depth of knowledge and accessibility are more important at the advanced levels.

The most widespread category of lemmas in the corpus is that of nouns (36.17%), followed by adjectives (16.32), verbs (13.55%), interlanguage expressions (7.48%), proper nouns (5.23%), adverbs (4.67%), multi-word lexical units (4.65%), geographical nouns (3.67%), foreign words (2.55%), numbers (1.41%), pronouns (1.28%), interjections (1.12%), conjunctions (1.06%), prepositions (0.80%) and articles (0.04%). An important result is that the category of interlanguage expressions is very relevant, because it is the specific mark of the LIPS corpus and demonstrates the relevance of this kind of annotation in a learner corpus. If I then consider the relation between tokens and grammatical categories, the main result is that verbs and nouns are the most widespread categories, highlighting the relevance of these conceptual categories in the learning process.

In figure 11 below, we can see how the distribution of the grammatical categories in every proficiency level does not change much, even though there are some relevant differences: adverbs increase at the advanced levels, multi-word lexical units and pronouns increase from level A2 onwards, verbs from level B1, interlanguage expressions progressively decrease after level B2, as do interjections after level B1. Function words enter and become stable in lexical competence quite early at beginner levels, content words progressively increase at the more advanced levels.

	A1	%	**A2**	%	**B1**	%	**B2**	%	**C1**	%	**C2**	%
Adjectives	45	10.84	124	8.34	1101	8.72	1278	8.69	1217	8.54	1619	8.51
Adverbs	26	6.27	133	8.95	1386	10.98	1775	12.07	1737	12.19	2144	11.27
Conjunctions	46	11.08	143	9.62	1344	10.64	1308	8.90	1326	9.31	1994	10.48
Articles	31	7.47	134	9.02	1109	8.78	1345	9.15	1202	8.44	1756	9.23
Foreign words	1	0.24	5	0.34	44	0.35	71	0.48	49	0.34	106	0.56
Interjections	24	5.78	102	6.86	583	4.62	631	4.29	693	4.87	724	3.81
Interlanguage expressions	4	0.96	16	1.08	157	1.24	133	0.90	121	0.85	123	0.65
Multi-word lexical units	2	0.48	25	1.68	202	1.60	231	1.57	258	1.81	310	1.63
Geographical nouns	21	5.06	29	1.95	229	1.81	311	2.12	112	0.79	137	0.72
Nouns	67	16.14	227	15.28	1867	14.79	2283	15.53	2133	14.97	2957	15.54
Proper nouns	1	0.24	16	0.94	117	0.93	84	0.57	54	0.38	61	0.32
Numbers	1	0.24	13	0.87	31	0.25	25	0.17	60	0.42	30	0.16
Prepositions	66	15.90	156	10.50	1232	9.76	1509	10.26	1424	10.00	2175	11.43
Pronouns	18	4.34	113	7.60	938	7.43	1043	7.09	1183	8.31	1487	7.82
Verbs	62	14.94	252	16.96	2287	18.11	2676	18.20	2675	18.78	3401	17.88

Figure 11 Grammatical categories in different proficiency levels

Figure 12 below shows the results of the groups based on the learning context. They are quite similar, even if there are some differences between learners exposed to spontaneous context and learners who have never been directly in contact with an Italian context.

	Outside of Italy	%	**Italy**	%	**Mixed**	%
Adjectives	2906	8.44	1798	9.16	680	8.04
Adverbs	3908	11.35	2403	12.25	890	10.53
Conjunctions	3446	10.01	1843	9.39	872	10.32
Articles	3198	9.29	1600	8.15	779	9.22
Foreign words	206	0.60	47	0.24	23	0.27
Interjections	1622	4.71	504	2.57	631	7.46
Interlanguage expressions	410	1.19	106	0.54	38	0.45
Multi-word lexical units	582	1.69	306	1.56	140	1.66
Geographical nouns	465	1.35	248	1.26	126	1.49
Nouns	5205	15.12	3089	15.74	1240	14.67
Proper nouns	175	0.51	113	0.58	43	0.51
Numbers	63	0.18	71	0.36	26	0.31
Prepositions	3627	10.54	2054	10.47	881	10.42
Pronouns	2521	7.32	1647	8.39	614	7.26
Verbs	6091	17.69	3792	19.33	1470	17.39

Figure 12 Grammatical categories in different learning contexts

In the "Outside of Italy" group there are more foreign words and interlanguage expressions than in the other groups, as its competence tends to deviate from standard Italian more than the competence of learners exposed to a more spontaneous input. Learners learning Italian in Italy have more opportunities to be exposed to native speakers' input, to assimilate this input and to make their competence come closer to the standard Italian of native speakers. Interjections are not very common in the group "Italy", and are more used in the other groups. This is maybe due to the different fluency that characterizes the spoken language of learners with different learning processes. Learners who do not live in contact with the natural context normally use more interjections to compensate for a lack of fluency and as a lexical strategy for filling lexical gaps.

Learners start to learn function words at the beginner levels, because this kind of word is very frequent in the learning input, especially in the spoken language. Learning input plays a crucial role in the lexical development and determines the sequence of acquisition of content and function words. Furthermore, as the function words are part of a closed class of words, their acquisition can become stable quite early in the learning process, while content words can always increase not only in second-language learning, but also in first-language learning.

Extrinsic measures of lexical richness: comparison with native speakers' vocabulary

Extrinsic measures of lexical richness have the advantage of taking into consideration not only the quantity of words a speaker uses, but also their quality, their frequency in native speakers' use and in input, and therefore their relevance for learners. They are based on the assumption that learners first learn the words most frequently used by native speakers and hence use these words more frequently. Beginner learners mostly use high frequency word, while advanced learners can use not only high frequency words, but also low frequency words. For Italian language there are no extrinsic measures of lexical richness as there are for English (Laufer 1995, 1998; Laufer & Nation 1995; Meara & Bell 2001), so I used the frequency list of native speakers and the Italian core vocabulary to compare first of all the lexical coverage of learners of Italian as L2.

I compared the lexical coverage of the first 2000 words of the frequency lists of mine corpus, of the entire LIPS corpus, of LIP (the native speakers' spoken frequency list), and of LIF (the native speakers' frequency list) (see figure 13).

Words	***Corpus***	**LIPS**	**LIP**	**LIF**
1-500	84.13%	82.24%	80.40%	78.07%
501-1000	90.09%	88.51%	85.99%	84.54%
1001-1500	92.50%	91.45%	89.07%	88.21%
1501-2000	95.30%	93.17%	91.07%	90.69%

Figure 13 Lexical coverage

The lexical coverage of the sub-corpus and of LIPS is smaller than the lexical coverage of both LIF and LIP, so the non-native lexicon is poorer than the native lexicon. Non-native speakers use the same number of tokens to produce more speech than the native speakers. If I consider the first 50 words of the frequency lists, it is very significant that the lexical coverage is almost the same for native and non-native speakers. High frequency words are the most used words both by native and non-native speakers, demonstrating the role of input, represented by native speakers' lexical uses, for the learning process.

I then compared the 100 most frequent words of mine corpus and of LIPS with LIP, examining above all the presence of grammatical categories in the three frequency lists (see figure 14 below).

	Noun	Verb	Adjective	Pronoun	Adverb	Conjunction	Preposition
Corpus	15	18	9	10	16	11	11
LIPS	12	17	10	12	16	10	11
LIP	6	23	7	16	19	16	8

Figure 14 Grammatical categories of the 100 most frequent words of the frequency lists

As the above figure shows, the main difference is the one between nouns and verbs. It is very meaningful that in the corpus and in the LIPS we have more nouns and fewer verbs, because this probably due to the fact that non-native speakers learn first nouns and then verbs, as in the L1 learning process[3]. Nouns are more linked to reality than verbs, which are more used in advanced learners' competence. In the other categories we have some differences, probably due to the acquisition sequences, but they are not so broad. These data confirm the role of learning input in the learning process of an L2, as the input of native speakers is a crucial factor in the developing of lexical competence for non-native speakers. The tops of the frequency lists of corpus, LIPS and LIP are very similar, presenting a lot of common aspects: the presence of high frequency nouns and verbs with a very general meaning, which are part of the Italian core vocabulary, many words and expressions typical of the spoken language such as interjections and other words like *sì* (yes), *no* (no), *allora* (then), *adesso* (now) and *bene* (well).

I lastly compared the corpus with VDB, the Italian core vocabulary. I compared it with VDB for each proficiency level and for each group of learners with different learning inputs. At level A1 67.5% of words belong to VDB, at level A2 65.2%, at level B1 58.77%, at level B2 58.16%, at level C1 60% and at level C2 60.78%. Core vocabulary decreases from the beginner levels to the intermediate levels and increases slightly at advanced levels. Learners learning Italian outside of Italy use 54.04% of core vocabulary words, learners learning Italian in Italy 58.8% and learners learning Italian in both contexts 64.63%. Learners more directly exposed to the input of the Italian context use more words from VDB, acquiring lexical uses more similar to those of native speakers and to the Italian core vocabulary. Compared with the other groups, learners from the group 'outside of Italy' use fewer words from VDB, not because they use many low frequency words, but because in their spoken production there are a lot of interlanguage expressions, foreign words, nouns referring to objects and places depending on the individual lexical learning process that cannot be compared to VDB.
VDB is divided in three sections based on the different frequency of words: fundamental lexicon (FO), high use lexicon (HUL) and high availability lexicon (HAL).

[3] On the relationship between L2 vocabulary learning and grammatical categories see Nation (1990: 48), Broeder *et al*. (1993), Bogaards (1994: 152), N. Ellis (1994: 251), Singleton (1999: 140) and Bettoni (2001: 72) for Italian as L2. Also in the L1 vocabulary learning many studies claim that children first learn nouns, followed by verbs and adjectives. On Italian as L1 see Caselli & Casadio (2002) and Laudanna & Voghera (2002). See also Aitchison (1994) on the distribution of the grammatical categories in the mental lexicon and de Groot (1993) on the effects of different kinds of words on the organisation of the mental lexicon.

I analysed the VDB words of the corpus to see the distribution of the three sections. 60.95% of corpus words belong to the fundamental lexicon, 28.97% are high use lexicon and 10.07% are high availability lexicon.

Figure 15 below shows the results of the comparison for each group of learners based on the learning context.

	Outside of Italy	**Italy**	**Mixed**
FL	66.88%	70.4%	76.62%
HUL	24.64%	22.99%	17.62%
HAL	8.46%	6.6%	5.75%

Figure 15 Core vocabulary in learners with different learning contexts

The mixed group is the one with the bigger difficulties in using the less frequent words from VDB. The 'Italy' and the 'outside of Italy' groups use lots of less frequent words from VDB, which is to say that in the learning process less frequent words are also very relevant for learners, who are induced to learn them in every learning context. Figure 16 presents the results for the group based on the proficiency level.

	A1	**A2**	**B1**	**B2**	**C1**	**C2**
FL	88.46%	86.6%	73.31%	71.64%	72.48%	73.43%
HUL	6.73%	9.97%	18.5%	21.32%	21.07%	21.43%
HAL	4.81%	3.44%	8.19%	7.04%	6.45%	5.15%

Figure 16 Core vocabulary at different proficiency levels

At beginner levels, namely A1 and A2, learners use more words of the fundamental lexicon than at other levels. At level B1 learners use a lot of high availability words and at level C2 they use many high use words. The distribution of the three divisions is quite irregular among the levels. This irregularity forces me to make only some generalizations about the development of lexical competence, observing the increasing presence of words of every section from the beginner levels to more advanced levels, confirming thus the relevance of VDB words in the vocabulary learning of Italian as L2.

Conclusions

The development of lexical competence both in natural and formal contexts is highly influenced by many factors, among which learning input is a fundamental one. In acquisition processes the size dimension develops more than in formal learning contexts, perhaps because the access to a broader input, such as the input of a natural context, fosters an increase in the quantity of words learners know. The size dimension develops especially at the B2-C1 levels, and then becomes less relevant in making the

difference between learners with different proficiency levels. What particularly differentiate more advanced from less advanced learners are depth of knowledge and accessibility.

Different learning inputs also determine different degrees of lexical richness. Different learning inputs greatly affect the development of lexical competence: learners acquiring Italian in Italy have a deeper knowledge of the Italian lexicon than learners learning Italian outside of Italy[4]. Looking at learners with different proficiency levels, there is an increase of lexical richness in the vocabulary of non-native speakers along the learning continuum, even though there is a decrease after a certain level due to the fact that to describe the advanced lexical competence it is also necessary to consider the depth of knowledge, lexical organization and the access to the lexicon.

The lexical uses of non-native speakers reflect the lexical habits and the lexical tendencies of native speakers. The vocabulary of non-native speakers is quite close to the lexicon of native speakers, especially when observing the most frequently used words. The most frequently used words in the native input are the words learned earlier, at the beginner levels.

The complexity and variability of learners' vocabulary are a mirror of the complexity and variability of the native uses of Italian, and of the impact of every single learning context and process. The great variability of spoken Italian generates phenomena of variation and irregularity in the interlanguage varieties and every single learning context can determine individual lexical development processes.

As far as the practical applications of mine corpus and of LIPS are concerned, the first is the realization of *DIS - Italian Dictionary of Uses for Foreigners*, at present at the planning stage, a monolingual dictionary aimed at serving both teachers of Italian as a second language and non-native learners. LIPS could also validate the CILS examination, helping with the selection of the texts and the exams for the proficiency evaluation. Finally, it could be a good starting point for the development of syllabuses and curricula, as well as for the production of didactic materials.

In the near future I want to study the sub-corpus and the LIPS, concentrating on the following aspects. The perspective of the research is to compare the acquisition models for Italian as a second language with the dimension of vocabulary, which has been little studied so far insofar as Italian as L2 is concerned. This is due to the fact that Italian research has concentrated more on morphological and syntactical aspects of learning Italian as a second language.

I would then like to compare both the corpus and LIPS with the receptive vocabulary texts of the CILS examinations of the same learners to see what differences there are between these two aspects of lexical competence. Finally, I would like to compare the oral exams with the written proficiency exams of the same examinees to see what differences there are between the oral and written productions of the same learners. The corpus LIPS is also available on the web-site www.parlaritaliano.it

[4] On the effects of the learning context on lexical development see Collentine (2004), who states that natural contexts affetct in positive way the lexical development compared to formal learning contexts, and for Italian L2 see also Bernini (2003) and Spreafico (2005).

References

Bagna, C. & Carloni, F. & Machetti, S. (2003). Il lessico del parlato degli stranieri in Italia. In Albano Leoni, F. Cutugno, F. Pettorino, M. Savy, R. (eds), *Il parlato italiano, Atti del Convegno nazionale di Napoli, 13-15 febbraio 2003*. Naples: M. D'Auria Editore.

Ball, F. (2001). Using corpora in language testing. *Research Notes*, 6: 6-8.

Ball, F. (2002). Developing a wordlist for BEC. *Research Notes*, 8: 10-13.

Barker, F. (2004). Using corpora in language testing. *Modern English Teacher*, 13, 2: 63-67.

Barni, M. & Gallina, F. (2008). Le parole degli stranieri: il LIPS, il primo lessico di frequenza dell'italiano parlato dagli stranieri. In Barni, M. Troncarelli, D. Bagna, C. (eds), *Lessico e apprendimenti. Il ruolo del lessico nella linguistica educativa*. Milan: Franco Angeli: 143-156.

Barni, M. & Gallina, F. (2009). Il corpus LIPS (Lessico dell'Italiano parlato da Stranieri): problemi di trattamento delle forme e di lemmatizzazione. In Andorno, C. Rastelli S. (eds), *Corpora di italiano L2. Tecnologie, metodi, spunti teorici.* Perugia: Guerra Edizioni: 139-151.

Bernini G. (2003). Come si imparano le parole. Osservazioni sull'acquisizione del lessico in L2. *Itals,* I, 2, 23-47.

Bettoni C. (2001), *Imparare un'altra lingua*. Roma – Bari: Laterza.

Bogaards P. (1994). *Le vocabulaire dans l'apprentissage des langues étrangeres*. Paris: Hatier Didier.

Broeder, P. & Extra, G. & van Hout, R. (1993). Richness and variety in the developing lexicon. In Perdue, C. (ed), *Adult language acquisition: cross-linguistic perspectives*. Cambridge: Cambridge University Press, 145-163.

Carloni, F. & Vedovelli, M. (2005). Il vocabolario di base dell'italiano degli stranieri. In De Mauro,T. & Chiari I. (eds), *Parole e numeri, Analisi quantitative dei fatti di lingua*. Rome: Aracne Editrice, 247-275.

Caselli, M.C. & Casadio, P. (2002). *Il primo vocabolario del bambino*. Milan: Franco Angeli.

Collentine J., (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies on Second Language Acquisition*, 26, 2, 227-248.

De Groot A.M.B., (1993). Word-type effects in bilingual processing tasks:support for a mixed-representational system. In Schreuder, R. Weltens, B. (eds.). *The bilingual lexicon*. Amsterdam – Philadelphia: John Benjamins Publishing, 27-51.

De Mauro, T. (1980). *Guida all'uso delle parole*. Rome: Editori Riuniti.

De Mauro, T. & Mancini, F. & Vedovelli, M. & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milan: ETASLIBRI.

Ellis, N.C. (ed), (1994). *Implicit and explicit learning of languages.* Edinburgh: Academic Press Ltd.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (1995), Modified oral input and the acquisition of word meanings. *Applied Linguistics*, 16, 4, 409-441.

Granger, S. & Wynne, M. (1999). Optimising measures of lexical variation in EFL learner corpora. In Kirk, J. (ed), *Corpora Galore*. Amsterdam – Atlanta: Rodopi, 249-257.

Haastrup, K. & Henriksen, B. (2000). Vocabulary acquisition: acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, 10, 2, 221-240.

Henriksen, B. (1999). Three dimensions of vocabulary development, *Studies on Second Language Acquisition*, 21, 2, 303-317.

Laudanna A., Voghera M., (2002). Nouns and verbs as grammatical categories in the lexicon. *Journal of Italian linguistics*, 14, 1, 9-26.

Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. In Eubank, L. Selinker, L. Sharwood Smih, M. (eds), *The current state of interlanguage: studies on honor of W.E. Rutherford*. Amsterdam – Philadelphia: John Benjamins Publisher, 265-272.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different?. *Applied linguistics*, 19, 2, 255-271.

Laufer, B. & Nation, P. I. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied linguistics, 16*, 3, 307-322.

Laufer B. & Paribakht, T.S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language learning*, 48, 365-391.

Meara, P. (1990). A note on passive vocabulary. *Second Language Research*, 6, 2, 150-155.

Meara, P. (1996). The dimension of lexical competence. In Brown, C. MlamKjaer, K. Williams, M. (eds), *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press, 33-53.

Meara, P. (1999). *The vocabulary knowledge framework*, www.lognostics.co.uk.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospects*, 16, 3, 5-19.

Melka, F. (1997), Receptive vs. productive aspects of vocabulary. In Schmitt, N. Mc Carthy, M. (eds), *Vocabulary. Description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 84-102.

Mondria, J.A. & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: what differences does it make? In Bogaards, P. Laufer, B. (eds), *Vocabulary in a Second Language*. Amsterdam – Philadelphia: John Benjamins Publishing, 79-100.

Nation, P. (2001). *Learning vocabulary in another language*, Cambridge: Cambridge University Press.

Palmberg, R. (1989). What makes a word English? Swedish speaker learners' feeling of "Englishness". *AILA Review*, 6, 47-55.

Read, J. (2004). Plumbing the depths: how should the construct of vocabulary knowledge be defined? In Bogaards, P. Laufer, B. (eds.), *Vocabulary in a Second Language*. Amsterdam – Philadelphia: John Benjamins Publisher, 209-227.

Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes*, 20, 12-16.

Schmid, H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Paper presented at the International Conference on New Methods in Language Processing.

Schmid, H. (1995). *Improvements in Part-of-Speech Tagging with an Application to German*. In Proceedings of the 14th International Conference on Computational Linguistics.

Schmitt, N. & Mc Carthy, M. (eds), (1997). *Vocabulary. Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schreuder, R. & Weltens, B. (eds), (1993). *The bilingual lexicon*. Amsterdam – Philadelphia: John Benjamins Publishing.

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Spreafico, L. (2005). Lo sviluppo lessicale di un apprendente di italiano L2. Problemi e metodi di analisi quantitativa. In Banti, G. & Marra, A. & Vineis, E. (eds). *Atti del 4° congresso di studi dell'Associazione Italiana di Linguistica Applicata, Modena 19-20 febbraio 2004*, Perugia: Guerra Edizioni, 241-257.

Vedovelli, M. (ed), (2005). *Manuale della certificazione dell'italiano L2*. Rome: Carocci.

Vedovelli, M. (2006). Il LIPS - Lessico di frequenza dell'Italiano Parlato dagli Stranieri. In Bardel, C. Nystedt, J. (ed), *Progetto Dizionario Italiano-Svedese. Atti del primo colloquio, Stoccolma, 10-12 febbraio 2005, Acta Universitatis Stockholmiensis 22*. Stockholm: Romanica Stockholmiensia, 55-78.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17, 1, 65-83.

Vermeer, A. (2004). The relationship between lexical richness and vocabulary size in Dutch L1 and L2 children. In Bogaards, P. Laufer, B. (eds), *Vocabulary in a Second Language*. Amsterdam – Philadelphia: John Benjamins Publisher, 173-189.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon. A depth of individual work knowledge model. *Studies on Second Language Acquisition*, 23, 1, 41-69.