# Recurrent Sequences in a Learner Corpus of Computer-mediated Communication

To Max and Francesco
for all your patience.

To my mum for being there,
And to my beloved Miranda,
who one day might forgive me for all
the time I did not spend with her.

# Contents

# Chapter One

# Introduction

The present research is an analysis of recurrent sequences of words in a corpus of learner English collected by means of computer-mediated communication, or CMC. Firstly, it is a contribution to studies of learner corpora, one of the most fruitful areas of learner English research over the last decades (Granger (1998, 2004b); Nesselhauf (2005); Campoy and Luzon (2007); Aijmer (2009a); Gilquin *et al.* (2008)) and secondly to phraseology, another field which has attracted a wealth of studies since the early 1990s. In addition, it introduces a study of learner English produced by means of computer-mediated communication, a medium which has been shown to have specific features across different languages and cultures (Herring, 1996, 2002; Macfadyen, 2006b; Thorne and Black, 2007; Herring, 2010).

In the literature, recurrent sequences of words have been given a variety of names, from *semi-preconstructed phrases* (Sinclair, 1991), to *lexical phrases* or *phrasal lexemes* (Nattinger and DeCarrico, 1992; Moon, 1998), *habitual collocations* (Fernando, 1996), *word combinations* (Cowie, 1998), *lexical bundles* (Biber *et al.*, 1999), *formulaic sequences* (Wray, 2000) and *phraseologisms* (Gries, 2008). Broadly speaking, and for the purpose of the present research, they can be defined as word sequences whose high frequency suggests speakers produce them as prefabricated chunks which carry out specific functions of discourse. Corpus-driven research into native-speaker English, such as Biber *et al.* (1999, 2004),

Biber and Barbieri (2007) and Biber (2009), indicates that frequently repeated word sequences show different structural and functional features depending on register and mode of communication.

The present thesis seeks to ascertain whether recurrent sequences of words exhibit the same features in a different mode of communication, CMC. Learners of English are increasingly using this CMC both in learning contexts and in their private lives. Computer-aided language learning (CALL) has been shifting its focus from the computer to communication, because CMC provides a real communicative context which learners can take advantage of and which they generally find more motivating.

Research has argued that recurrent word sequences play a major role in learner English 'idiomaticity', that is the ability to convey meaning using combinations that are grammatically correct as well as native-like, it is also believed that the study of sequences provides insights into the processing of language by learners. Psycholinguistic research, in fact, has repeatedly sought evidence that sequences of words are produced holistically. Before the advent of learner corpora, learner English used to be described as constructed of individual 'building blocks' of meaning rather than prefabricated expressions (Kjellmer, 1994). However, in more recent studies, evidence shows that recurrent sequences, rather than individual words, are the 'basic building blocks' learners employ to construct their discourse (De Cock, 2004). At the same time, they are called 'stumbling blocks' for learners; most studies claim that learners use fewer sequences compared to native-speakers, that they employ them with different functions, regardless of the register they are using for communication. In native-speaker English, prefabricated language has been proven to be influenced by register and mode of production. Similarly, it is expected that learner English will also be influenced by register. The use of recurrent sequences in learner asynchronous chats, therefore, will provide further insights into learner English differences across registers and modes of production.

The present study, whose research questions and aims are related in detail in Section 1.1, is based on a corpus of learner CMC English collected and assembled for the purpose and devised to be comparable to existing learner corpora. The collection of data through asynchronous chats introduces a new variable that is important for the understanding of the relationship of sequences to means of communication in learner English. Language produced in computer chats resembles face-to-face conversation in terms of its pace and for its imitation of the phonetic qualities of speech by means of typographic conventions. At the same time, however, it is distanced from the speaker by the electronic medium and appears as written text on the computer screen while it is being produced. This makes it available to inspection from the outside, something that is not possible while producing spoken discourse. This feature, among others, places CMC midway between speech and writing and provides an opportunity for learners not to feel the pressure of language processing, and monitor their output at the same time.

## 1.1 Aims of the Research

The aim of the present research is to investigate recurrent sequences of words in learner English from a corpus of asynchronous chats collected in an academic context in order to answer the following research questions, collected into three broad areas:

1. What recurrent word sequences can be found in advanced learner English collected by means of the computer? What are the structures and functions of these sequences? What features prevail? Are they more similar to speech or writing?

2. Are recurrent sequences influenced by computer-mediated communication? Do they differ from recurrent sequences in comparable corpora of learner writing and learner speech? Are the same sequences used in learner corpora? Does each mode of communication employ specific types of sequences?

What can recurrent sequences tell us about adaptation to register in learner English?

3. What can recurrent sequences of words tell research about how the learners process the language they are producing?

The answers to these questions are sought by means of quantitative and qualitative analyses of a specially compiled corpus and findings are discussed in the light of corpus-based research on native-speaker and learner English. For the purpose of the present study, recurrent sequences of words, which can also be called N-grams, were extracted by means of the N-gram function of a concordancer and subsequently analysed in terms of functions, structures and register differences, analogously to the work carried out by Biber *et al.* (1999) in the identification and study of lexical bundles for the *Longman Grammar of Spoken and Written English*. Biber *et al.*'s terminology is also employed with the use of register to indicate text-type, with reference to the ones that were found to be the furthest apart in native-speaker English: conversation and academic writing. Within the descriptive approach exemplified by Biber *et al.*'s grammar and Biber and colleagues' successive research on learner bundles, the present study attempts to provide an accurate description of the linguistic features, structures, functions and specificities of the learner English sequences extracted from the corpus.

In terms of mode of production, previous research on corpora of learner English has focussed mainly on written texts, essays and exams, and spoken, teacher-led interviews. Since there is, to date, no publicly available corpus of learner CMC English, the current study assembled a corpus which was designed to be comparable to the most influential learner English corpus, ICLE (2002) and, therefore, to enable cross-corpus comparisons and generalisability of findings. It is argued that the computer chats used for data collection in the present research provide an additional mode of communication which can be compared to the written and the spoken ones. For learners, communicating in English by means of the computer is a real-life activity, characterised by lower levels of anxiety compared to

spoken interviews, and higher levels of motivation compared to written essays (as discussed in Sections 2.6.3 and 3.3.5). Moreover, chats have the further advantage of giving learners extended processing time and extra time for editing. As a result, CMC brings with it processing and motivational advantages, as well as mode-specific features, which should be reflected in the use of sequences and provide further insights into the features of learner English and of learner language processing.

## 1.2 Context of the Research

The study of recurrent sequences of words in learner English by means of a corpus of learner chats encompasses different areas of linguistic research, namely, research on phraseology, on learner corpora and on computer-mediated communication. The following paragraphs place the present research into context in order to show the research gap it addresses and tries to bridge.

Over the past thirty years, linguistics has devoted considerable attention to the study of the phraseology of words in context, both in native-speaker and in learner English (Fernando, 1996; Cowie, 1998; Moon, 1998; Wray, 2000; Schmitt, 2004; Nesselhauf, 2005; Wray, 2005; Meunier and Granger, 2008; Wulff, 2008; Wood, 2010b). The interest in word combinations mainly stems from John Sinclair's frequency-based approach to collocation and phraseology. Sinclair (1991, 1996) and Sinclair and Carter (2004) claimed that collocation should be at the centre of corpus linguistics analysis, and developed the fundamental notion that semi-preconstructed phrases constitute 'single choices' for the language user. According to Sinclair, fixed and semi-fixed expressions have a meaning as wholes for the language user and, therefore, they are mentally stored as single units (Sinclair, 1991:110).

Behind this crucial shift in perspective lay the collaboration of linguistics with computer science. As computers made it possible to process large amounts of text in a wide range of text types and analyse them using software tools that could

automatically extract frequently occurring sequences of words, more and more linguists were attracted to the study of word combinations. Some took a corpus-based approach, that is, they used corpora as evidence for linguistic theories, others a more radical corpus-driven approach (Francis, 1993; Tognini-Bonelli, 2001), that is, they analysed the corpora by means statistical and computational counts and then analysed the automatically retrieved data. Both approaches have revealed new facts about the language, which have greatly influenced the modern approach to the study of language in context.

From a psycholinguistic perspective, research has investigated processing and storage of collocation, word combinations and idioms in native and non-native speakers of English using both corpus methodology and experimental methods (Durrant and Doherty, 2010; Siyanova-Chanturia *et al.*, 2011; Tremblay *et al.*, 2011). Although studies seeking evidence of the theory of holistic storage of word combinations have not yet yielded conclusive results, scholars tend to agree that recurrent sequences of words are used to reduce language processing effort. In this respect, a description of the use of recurrent sequences of words in a corpus of learner CMC may give further insights into the processing of language chunks by learners of English.

Corpora of learner English have been collected and analysed since the 1990s. ICLE, the International Corpus of Learner English (2002), was the first major corpus of learner English writing collected from multiple language backgrounds in a chiefly European EFL context. A large part of the most influential studies of learner language is based on this corpus and, more recently, on its sister corpus, the Louvain International Database of Spoken Interlanguage, or LINDSEI (2010). Over the past ten years, other learner corpora have been assembled and analysed with a variety of methodologies. Corpora of learner CMC English have only recently attracted the attention of research in EFL and some scholars have started collecting data from electronic communication, such as Foss (2009) in Japan and Ädel (2011) in Sweden.

In the analysis of learner corpora, most existing studies use a comparative approach with native-speaker corpora (De Cock, 2004; Aijmer, 2004; Nesselhauf, 2005; Aijmer, 2009b), or contrastive interlanguage analysis (Gilquin *et al.*, 2008). The first approach is undoubtedly the most appropriate for the identification of overuse, underuse and misuse of specific structures or lexical items (Granger and Rayson, 1998), while the second approach has been employed for the study of transfer in the context of error analysis (Granger, 1996 and Gilquin, 2000 *inter alia*).

However, the study of recurrent sequences in learner English poses a number of problems. Firstly, there is no ready-made list of native-speaker sequences that can be used for extraction and analysis. Secondly, as argued by current research into the social and cultural implications of the use of recurrent sequences of words (Mair *et al.*, 2000; Skandera, 2007; Wierzbicka, 2009; Schauer, 2009), production of native-like sequences is strictly connected to belonging to the speech community that uses them, which is not the case when L2 acquisition happens in EFL countries. The notion of belonging, or aspiring to belong, to a community of speakers, or a community of practice, has been proven to be key to the production of native-like sequences in academic writing (Hyland and Milton, 1997; Hyland, 2008a) and it is believed it has an influence in the production of sequences in other registers and modes of communication. In this respect, the analysis of recurrent sequences in learner CMC English may provide corroborating evidence of the relationship between community of practice and register-appropriate sequences.

According to research, learners variably rely on prefabricated formulae and compositionality at different stages of second language development (Wray and Perkins, 2000) and scholars have suggested that learner language is produced by habit and repetition as much as by compositional strategies (De Cock, 2004). Moreover, learner English corpora have shown that learners employ spoken English features in academic writing and expressions which are more typical of written English in spoken interaction (De Cock *et al.*, 1998; Gilquin and Paquot,

2008). Overall, studies of learner corpora present heterogeneous, at times contrasting, views of learner English at advanced levels of competence, and there have been no attempts at providing a general description of learner English across registers.

In the relatively new research field of computer-mediated communication, studies have revealed that CMC English has acquired unique linguistic and typographic features (Herring, 1996; Murray, 2000; Herring, 2002; Thorne and Black, 2007; Herring, 2010). CMC has repeatedly been described as speech written down, but it has been proved that it shows features of different registers, depending on communication mode and participants. All over the world, learners of English are using CMC both as a means of language learning (CALL) and as a global means of communication. Network-based Language Teaching (NBLT) research has shown that CMC provides a real communicative context in which learners can practice the language with the pace of conversation, but with less pressure and more opportunities to monitor their own output.

CMC is a new mode of communication which has only recently attracted the attention of learner corpus research, and no study has yet compared recurrent sequences of words produced by learners across different registers and modes of communication. This research path is deemed to be interesting for two main reasons: firstly, it gives a unique opportunity to compare learner English across registers without referring to the largely unattainable native-speaker model. Secondly, it may provide further insights into learners' processing of the language in a new mode of communication which produces less anxiety than speech, and is more involving and closer to learners' real life experiences than academic writing.

## 1.3 Corpus as Methodology

The analysis of learner English presented here is grounded in corpus methodology. Within this framework, the data is extracted automatically by means of a concordancer. Quantitative findings are then compared across learner corpora

which have been subject to data extraction by means of the same software tool.

The major advantage of using a corpus to answer research questions on recurrent sequences of words in learner language is that a computerised corpus enables the researcher to uncover features of language use that may escape intuition. As a consequence, since the advent of corpus linguistics, even researchers who work deductively tend to adopt corpora to look for evidence to support their theories.

According to McEnery and Hardie (2012), another main strength of corpus linguistics is that it 'allows access to reliable information regarding frequency', a type of information that would be extremely arduous to collect using traditional techniques. Frequency findings are the basis of modern lexicography and they have had an enormous impact on the compilation of learner dictionaries and learner materials in general. They have also informed descriptive grammars like the *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999).

In terms of learner language, the collection of a corpus represents a unique opportunity to implement a bottom-up approach on large quantities of learner output and encourages new perspectives on learner language in general, which 'might even challenge some of the most-deeply rooted ideas about learner language' (Granger, 2004a). The collection of a new type of corpus of learner language, in fact, provides an opportunity to look at learner English with new eyes, from a different perspective.

The corpus compiled for the present research collects learners' asynchronous chats produced in an EFL context. It is best defined as an opportunistic corpus, that is, a corpus containing 'the data that it was possible to gather for a specific task' (McEnery and Hardie, 2012:11). Indeed, due to practical restrictions, learner corpora generally belong to this type. Therefore, research using corpora will always be able to report on the language a learner produces while carrying out a specific task in more or less controlled conditions, but not whether the learner possesses a specific lexical item, phrase or structure in their knowledge.

Principles and best practices for learner corpus collection are discussed both

in Tono (2003) and in (McEnery and Hardie, 2012). It is argued that detailed corpus design is key to controlling the variables that affect language production by learners. As a matter of fact, controlling variables ensures that the ensuing corpus is a suitable sample of learner language and responds to the criteria of accountability, falsifiability and replicability necessary to make any results derived from its analysis stand on solid ground.

As Granger (2004:126) points out, compared to other corpus data, learner corpus data is characterised by extreme variability. For this reason, the principles behind learner selection and data collection for the present research study closely follow those employed by Granger for ICLE (2002). In the present research study, accurate control of variables and comparability in design make cross-corpus comparisons possible and generalisations on learner English across modes of communication more reliable. In addition, the bottom-up approach employed for the identification of recurrent sequences of words, automatic extraction of N-grams, normalisation of frequencies and analysis of concordance lines, was applied to all the corpora, which enhanced the comparability even further.

Corpus methodologies have also had their detractors, and a frequent criticism that has been levelled at language description through computer corpora is that corpora 'allow us to observe language, but they are not language itself' (McEnery and Hardie, 2012:26). In other words, corpus data is performance data connected to a specific place and time and, like experimental data, it is indirect evidence for internal linguistic competence. An increasing number of scholars has criticised corpus-driven methodology in the sense described by Francis (1993) and Tognini-Bonelli (2001). The distinction between corpus-based and corpus-driven analysis was particularly emphasised at the beginning of corpus linguistics, but it has more recently come under serious criticism, since it is well-nigh impossible to approach data without any theory or categorisation in mind (Stubbs, in press).

## 1.4 Overview of the Research

The present research study starts with the review of the relevant literature in Chapter Two. Section 2.1 deals with the notion of phraseology and its development from the study of idioms and fixed formulae to the analysis of recurrent sequences of words and lexical bundles. Section 2.2 summarises research regarding recurrent sequences in terms of processing and storage in the mental lexicon. Section 2.3, instead, gives a brief account of research strands dealing with the social and cultural implications of phraseology and its use, while Section 2.4 concentrates on recurrent sequences as marking features of two different registers: academic writing and speech. Studies of recurrent sequences and second language acquisition are reviewed in Section 2.5. The chapter closes with an overview of the linguistic features of text-based computer-mediated communication and of studies of learner language produced by means of CMC.

Chapter Three is an account of the design, collection and building of the Learner Chat Corpus (or LCC). Sections 3.1 and 3.2 review existing learner corpora of both writing and speech collected in various learning contexts throughout the world, focussing on the International Corpus of Learner English (ICLE), whose design was followed closely in the learner English corpus collected for the present work. The following sections of the chapter report on the data collection and corpus building phases, and provide a detailed description of the features of the corpus and of the learners under observation.

Chapter Four delves into the LCC with quantitative and qualitative analyses of the recurrent sequences automatically extracted by means of concordancer. Section 4.1 presents, analyses and discusses the frequency data regarding 2- to 6-word sequences in order to provide preliminary answers to the research questions about frequent word combinations in learner CMC English. In Section 4.2, the most frequent 3-word sequences from the corpus are classified in terms of structures and functions with the aim of uncovering the features of the learners' recurrent sequences.

Chapter Five compare the recurrent learner sequences in LCC with those in ICLE_IT, a corpus of essays and exam papers, and LINDSEI_IT, a corpus of oral interviews, both by Italian L1 speakers. Learners' most frequent sequences in these two registers are analysed and classified in terms of structures and function in Section 5.2. Section 5.3 is a qualitative cross-corpus comparison of two specific sequences: the most frequent sequences including the verb *think* and the quantifier *a lot of*. These sequences were found to appear among the most frequent in all the corpora and are the subject of in-depth analysis and comparison.

Chapter Six draws the conclusions of the present study. The findings are summarised and discussed in the light of current research on learner English sequences across registers and on learner language processing. The limitations of the study are also pointed out, as well as directions for further study.

# Chapter Two

# Literature Review

This chapter reviews studies, concepts and theories that are relevant to the present research. Sections 2.1, 2.2 and 2.3 are an account of research on the recurrent sequences of words in English linguistics; Section 2.4 addresses features of formulaicity in native-speaker English writing and speech; and Section 2.5 is a review of studies of recurrent sequences in learner English writing and speech. The chapter ends with a review of scholarly research on computer-mediated communication in English by native and non-native speakers in Section 2.6.

The main questions addressed by research on recurrent sequences concern their definition and categorisation, their processing and storage in the mental lexicon, their connection with social and cultural factors and the differences in usage between spoken and written registers. The different studies attempting to provide answers to these key questions constitute the backbone of the present literature review and they provide a point of departure for the examination of recurrent sequences in learner English.

Although the interest in sequences in English predates the creation of large corpora, early studies, which were mainly based on smaller data samples or linguists' intuitions, are mentioned in Section 2.1, but are not presented or discussed in great detail. It was considered more relevant for the present study to focus on more recent findings by means of corpus methodology, experimental studies, or a mix of the two methods. The continuing discovery of facts about the phraseo-

logical quality of written and spoken English is such that the following sections can only attempt to do justice to the wealth of publications and findings in areas as diverse as semantics, pragmatics, natural language processing, language acquisition and psycholinguistics. Instead, the review of the scholarly literature on recurrent sequences in English will report, in some detail, on studies that are comparable to the present research in terms of type of data, methodology or findings.

## 2.1 Definitions and Categorisations: from Idioms to Recurrent Sequences

In the 1930s, a few English linguists, notably Palmer, Hornby and Firth, had recognised the importance of phraseology in English. However, the generativist view of language was largely dominant at the time, and mainstream linguistics did not study phraseology as a unified phenomenon. According to Stubbs's (2007:89) account of phraseology within corpus linguistics, phraseological expressions 'had been seen by most linguists only as an unrelated collection of oddities'.

The pervasive presence of recurrent word combinations in language was the subject of a seminal paper by Pawley and Syder (1983), which speculated about the nature of native-like fluency. They argued that recurrent sequences constituted the basis of fluency in native-speakers' conversational talk and estimated that mature speakers of English know 'hundreds of thousands' of institutionalized sentence stems, which are largely fixed and used for culturally recognized concepts and speech acts (Pawley and Syder, 1983:191).

To linguists in the 1980s, the tendency of speakers to repeatedly use combinations of words represented a puzzle. Why did speakers prefer institutionalised expressions to other grammatically possible alternatives even when they were longer and syntactically more complex? Why was the creative potential of syntactic rules not used to its full potential? Why was generative creativity con-

sidered unidiomatic, odd, or foreign? These were the questions behind Pawley and Syder's article. Analysing different ways of telling the time, they noted that institutionalised sentence stems had:

> 'a grammar that is unique in that they are subject to an idiosyncratic range of phrase structure and transformational restrictions; that is to say, by applying generally productive rules to these units one may produce an utterance that is grammatical but unnatural or highly marked.' (Pawley and Syder, 1983:192)

Therefore, formulaic sequences existed '...somewhere between the traditional poles of lexicon and syntax...' (Nattinger and DeCarrico, 1992:1) and, consequently, could not be explained by relying exclusively on one or the other of these two poles.

Before the advent of corpus linguistics and usage-based descriptions of English, studies of sequences attracted the attention of lexicographers and language teachers, who focussed mainly on *idioms*, or *idiomatic expressions*, or multi-word sequences that operated as single units, with varying degrees of semantic transparency. Later on, when quantitative studies based on corpora provided new insights into their use, some scholars questioned their usefulness to learners of English as a second language.

J.M.H. Sinclair's work on large collections of texts was instrumental in shifting scholarly interest in semantic transparency, opacity and fixedness of form to recurrence, co-occurrence (continuous or discontinuous) and co-selection. Sinclair's *idiom principle* stated that:

> 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments.' (Sinclair, 1991:110)

This principle introduced the key concept of 'single choice', which turned out to be a fundamental contribution to the study of phraseology in the English

language. Sinclair used a bottom-up approach to identify recurrent sequences and discovered that frequent combinations found by means of concordances and distributional measures did not fit predefined categories, but could contribute to shedding light on the phenomenon of native-like selection which linguists had found so puzzling.

Frequency and co-selection had fallen out of favour in post-behaviourist linguistic theory (see Ellis, 2002), but Sinclair considered frequency a central concept to language, because frequent words are not independent of the context. According to Sinclair, it is the phrase that carries the meaning. For example, the fact that the definite article *the* occurs in almost every line of printed English indicates that *the* plays a role in the composition of recurrent phrases (Sinclair, 1999:157) and contributes to their meaning. The linguist's task, therefore, is the identification and study of the meaning and functions of those phrases to find out facts about the speakers and the language they speak.

The initial emphasis on the study of relatively rare expressions characterised by fixedness and semantic non-compositionality limited, to a certain extent, the research on the phraseological tendency of the English language. Subsequent studies considerably widened the definition of sequences (see Aijmer, 1996; Fernando, 1996; Moon, 1998; Schmitt and McCarthy, 1998; Cowie, 1998) and phraseology came to be considered a continuum from pure idioms to free combinations. Fernando's (1996:3) descriptive account of idioms and idiomaticity in the English language is a clear example of how idiomaticity became more and more an umbrella term including a variety of linguistic phenomena.

In the same period, studies of phraseology based on corpus data started to be published. Among them was Moon's (1998) corpus-based account of fixed expressions and idioms (FEIs) in English, based on the Oxford Hector Pilot Corpus[1]. Following Halliday's (1978) model, she classified FEIs into ideational and

[1]The Oxford Hector Pilot Corpus was a joint Oxford University Press and Digital project which took place in the early 1990s. It developed a 20-million word corpus of spoken and written English, which served as a pilot for the British National Corpus.

interpersonal. After a corpus-based account of FEIs in the English language, she concluded that 'existing models and descriptions need to be revised in the light of emerging corpus evidence, in particular with respect to form and variation.' (Moon, 1998:309). Moon's study was among the first to employ a corpus and it brought to light the distribution of FEIs in different genres. She observed that 'the density of metaphors and proverbs seems to be greater in journalism than other text types, and pure idioms seem to be less common in spoken interaction than is often thought.' (*Ibid.*:309), with individual items showing preference for individual genres. In terms of patterning and variation, the corpus analysed provided evidence of the instability of the form of FEIs, undermining the widely accepted criterion of fixedness of form and showing that examples drawn from actual texts demonstrate creative manipulation of language. Moon concludes that descriptions of FEIs did not account adequately for their characteristics as observed in corpora, especially in terms of their discourse functions. Most importantly to the present review of definitions and categorisations of multi-word sequences, she points out that 'FEIs represent real chunks of language which do not conform neatly to abstract categories' (*Ibid.*:310).

Similar findings were reported in Aijmer (1996). From a discourse analysis perspective, Aijmer investigated the pragmatic function of conversational routines in speech and found that, although they had some common formal features with fixed expressions and idioms, they were also characterised by a great deal of formal variation. Analysing the various multi-word expressions used to thank, apologise, request and offer, Aijmer distinguished 'different degrees of lexicalisation on a scale of frozenness' and that there was no fixed cut-off point between different types of phrases (*Ibid.*:10).

Altenberg (1998) was also among the first scholars to employ a corpus (the London-Lund Corpus of Spoken English) to identify recurrent strings of words occurring more than once in identical form. He defined them 'preferred ways of saying things' and found that most of the strings were semantically fairly

transparent and could be formed by productive grammatical rules. Even though he questioned the assumption that these strings are stored and retrieved as wholes from the mental lexicon, he highlighted the fact that they are conventionally used by many speakers for specific pragmatic functions.

Clearly, the categorisation impetus of the traditional lexico-grammatical approach to phraseology was waning as scholars realised that it was a phenomenon of discourse as well as lexicon and grammar. As scholars had come to realise that it was not possible to separate form from function, they created categories that were more and more inclusive and therefore less and less agreed upon. The multiplicity of approaches and the lack of a common terminology, 'with different terms covering the same units and the same terms used to denote quite different units' (Granger and Paquot, 2008:28) is summarised in Figure 2.1, compiled by Wray (2002:9), showing the 'plethora of terms' used for formulaicity.

Figure 2.1: Terms for formulaicity from Wray (2002)

amalgams – automatic – chunks – clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[dioms] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units

Partly due to this terminological confusion, later studies of phraseology focussed more on attempts to integrate findings on word sequences with current linguistic theories and on investigations into possible practical applications of findings about phraseology in language teaching (see Cowie, 1998, Schmitt and McCarthy, 1998 and Schmitt, 2000).

Among the studies relying on corpus data, Biber *et al.* (1999) investigated recurrent sequences for the *Longman Grammar of Spoken and Written English* (*LGSWE*), a descriptive grammar of American and British English. Using the methodology of automatic extraction, Biber *et al.* identified *lexical bundles*, i.e. word forms that most commonly co-occur in longer sequences (*Ibid.*:989) across a variety of texts. Lexical bundles are interesting units because, even though they are often structurally incomplete, they were found to be classifiable into basic types. Their common feature is that they are typically used as building blocks: in conversation they introduce verbal and clausal structural units, and extended noun or prepositional phrases in academic prose (*Ibid.*:992).

Findings from corpus linguistics have considerably broadened the definition of the phraseological unit to the point that Gries (2008:5) sums it up as 'the co-occurrence of a form or a lemma of a lexical item and any other kind of linguistic element'. Such a definition has made their identification even more elusive.

Reliance on quantitative data has sometimes been criticised; however, we can agree with Stubbs (2002) that actual frequencies of occurrence are a key to the typical, conventionalised ways of expressing meanings and functions. Their relevance in current linguistic studies is recognised, such that they have come to constitute the core of recent theories of grammar such as cognitive and construction grammar, and are said to influence connectivist and emergentist theories of language acquisition (see Conklin and Schmitt (2008); Gries, 2008; Larsen-Freeman and Cameron, 2008; Ellis and Larsen-Freeman, 2009).

## 2.2 Storage and Processing of Recurrent Sequences

The second main strand of interest in recurrent sequences of words is concerned with finding evidence of how they are stored in the mind and with their role in fluent speech production. According to Wray and Perkins (2000), a formulaic

sequence is

> 'a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.' (Wray and Perkins, 2000:1)

This definition is a clear move away from issues of semantic noncompositionality and unanalysability of units, grammatical anomalousness and fixedness, in order to focus on the psycholinguistic and cognitive aspects of unit memorisation and storage.

Experimental studies on phraseological units are not as numerous as corpus-based studies, as the various methods of collecting evidence of the psycholinguistic reality and holistic storage of recurrent sequences have yielded contrasting results. The methods used include grammaticality judgements and intonational contour (Ellis, 2002), eye-tracking (Siyanova-Chanturia *et al.*, 2011 and Columbus, 2010), reaction times (Durrant and Doherty, 2010), self-paced reading (Tremblay *et al.*, 2011). Despite this variety of methods, scholars seem to agree that we still lack adequate experimental methods to investigate how formulaicity is actually processed in the human brain (Biber and Barbieri, 2007).

There are three main assumptions held by linguistic research regarding the psycholinguistic reality of recurrent sequences: firstly, sequences are stored and retrieved holistically and not analytically; secondly, speakers employ them to ease the process of language production; and, thirdly, they are psychologically salient, i.e. speakers and writers realise they are to be considered whole expressions.

According to cognitive linguistics and psycholinguistics, input frequency is at the basis of language acquisition. Ellis (2002) states that speakers learn common sequences of words as they learn intonation units, i.e. as a result of frequency. By means of an experimental study of reaction times, Durrant and Doherty (2010) found a strong overlap between the knowledge of sequences by native-speakers and

corpus frequencies and concluded that high frequency of occurrence in corpora is a reliable indicator of psychological priming between words.

In another study, Tremblay *et al.* (2011) conducted experiments involving self-paced reading and word and sentence recall to assess the processing advantages of lexical bundles. They saw that lexical bundles were connected to faster reading times and higher rates of recall at a later time. This was considered further evidence that native-speakers register frequency differences and produce language accordingly, and, therefore, recurrent sequences of words are not mere computational accidents.

In a recent experiment, Siyanova-Chanturia *et al.* (2011) used eye-tracking to assess processing advantage for idioms over novel phrases (e.g. *at the end of the day* versus *at the end of the war*) and found consistent positive evidence for native-speakers, who exhibited fewer and shorter eye fixations when reading idioms. Using the same technology, Columbus (2010) found that even non-semantic sequences such as *at the end of the* are processed faster than compositional ones such as *to eat a sandwich*.

All the studies mentioned above, however, focussed on the perception of written sequences and were produced under experimental conditions, which did not involve language production in natural settings. Lin (2012), for instance, argues there is still a lack of evidence of holistic storage and that the missing piece of the puzzle might be located in prosody, that is in the study of pauses and intonation. Scholars have repeatedly sought evidence of holistic storage and processing in phonological coherence. Erman (2007), for example, looked for support of the hypothesis of holistic storage in pause distribution and duration. Phonological studies state that certain types of pauses reflect processing effort, that is retrieval from the mental lexicon. Analysing the Bergen Corpus of London Teenager Language (COLT) and the London-Lund Corpus of Spoken English (LLC), Erman found that pausing in the middle of prefabs does not occur in conventionalised sequences, while it may occur in other types of sequences.

Results are still far from being consistent, however, since Lin's (2009, 2010 and 2012) studies of intonation boundaries indicate that there is no complete matching of phraseological unit and intonation unit boundaries for all types of sequences. Sentence builders and semantically transparent collocations, for instance, tend not to correspond to single intonation units. Research on the effects of holistic storage and processing on intonation also pointed out that the length of sequences might have an influence which has not yet been fully taken into account. It could be argued that analyses of the recordings of utterances have the inevitable drawbacks of all indirect methods: fluent and efficient production is not equal to storage in the mental lexicon.

After analysing the data collected through an online sentence production test featuring the verbs *give* and *take* and comparing it to two present-day corpora of American English, Gilquin (2005) concluded that multi-word sequences may not be stored holistically. Based on the claim by cognitive grammar that the more often a particular symbolic unit is encountered in real speech or writing, the more entrenched this symbolic unit becomes in the linguistic system and the more automatically the unit is accessed, without analysing its internal structure (Gries, 2008), Gilquin set out to discover whether 'entrenched sequences' still bear some connection with their component words. Data from her study suggested that there 'might still be some association between a multi-word sequence and its components, even in the case of a recurrent multi-word sequence' (*Ibid.*:168). However, Gilquin's assumption is not universally shared. Tremblay *et al.* (2011), for example, suggest that 'it is also possible that the mental lexicon stores the continuous and discontinuous chunks contained within compositional multiword sequences...' (*Ibid.*:596) and it might be argued that, as a consequence of this, the component words of sequences may always be clear in the speakers' minds. Along similar lines, Bybee's (2010) 'Linear Fusion Hypothesis', argues that when two or more words are often used together, they 'fire together', meaning that they are activated together in the brain. A comparable concept is the so-called 'fusion

effect': when units are no longer analysed they fuse together first in pronunciation (as in the phonological reduction of *would have*, which becomes *would of*) and then in orthography (as in the connector *nevertheless*).

De Cock (2004:225) observes that language users are 'creatures of habit' because they tend to use conventional rather than novel ways of expressing themselves. In other words, sequences are used and reused by speakers because they were used frequently and successfully by a large number of speakers before them (Stubbs, 2009). The reasons for this linguistic behaviour are said to be linked to processing advantages. The original hypothesis, put forward, among others, by Pawley and Syder (1983), was that the mind finds it easier and quicker to retrieve ready-made chunks rather than generate the language through grammatical or syntactic rules. This strategy enables speakers to save processing effort while achieving their interactional goals (Wray, 2000), that is to say: formulae are a 'linguistic solution to a nonlinguistic problem...' (Wray, 2002:100). In spite of the lack of direct evidence discussed above, this concept is intuitively so logical that it has been taken at face value by most literature on phraseology. Spoken language, with its typical dysfluencies, has frequently been considered evidence of the holistic processing of chunks of language. Its high formulaic repetitiveness is explained by the influence of communicative pressure and context (Wray, 2008). These claims are supported by studies of conventionalised varieties of speech such as Kuiper's (2004), whose evidence from professional auctioneers and sports commentators' discourse demonstrates that under heavy time constraints speech becomes more formulaic.

The literature on formulaic sequences includes analyses of the processing of metaphors and classical, non-compositional idioms. Conklin and Schmitt (2008) used formulaic sequences with idiosyncratic meaning in a self-paced reading task in order to verify the hypothesis that reading times are quicker for formulaic rather than literal meanings of sequences. An interesting finding of this study was that formulaic sequences were read with the same speed, regardless of context

supporting their literal or metaphorical interpretation. Once again, it should be observed, however, that ease of processing does not equate with holistic storage.

Summing up, recent investigations into the processing of recurrent sequences have provided evidence that native-speakers seem to register frequency differences and produce language accordingly. However, as is pointed out in Tremblay *et al.* (2011), we currently do not have 'a clear picture of how sentences are actually assembled...' (*Ibid.*:596). There is no conclusive evidence that sequences are stored and processed as single units; they may simply be computed more efficiently in the speaker's mind. This does not detract from sequences relieving online pressure in language production and is considered proof that they actually leave traces in the brain.

## 2.3 Social and Cultural Implications of Recurrent Sequences

In an extensive review of the history of formulaic language research, (Pawley, 2007:22) notes that 'speaking a language idiomatically is a matter of conforming to established ways of saying things'. In other words, members of the community share 'subject matter codes' (*Ibid.*:23), that is, conventions specifying what can be said about a topic, when and why, but also with which word sequences. Content, form, context and purpose are all bound together and although there may be countless grammatically correct ways of expressing an idea, most of them will not be used by native-speakers of any given language.

Speech behaviour is shaped by a number of factors including experiences of being in the company of other members of the community and conventions governing face-to-face meetings. Native-speakers acquire speech competence largely out of awareness, which is exemplified by the fact that conventions tend to be noticed when they are violated by a newcomer to the community (Lee, 2007:471). Ellis (2002:157) argues that native-like competence, fluency, and idiomaticity re-

quire 'an awful lot of figuring out which words go together'. This 'figuring out' takes place throughout the long and complex process of language acquisition, usually starting in childhood and continuing for the speaker's entire lifetime. In addition to this, studies indicate that throughout adult life the store of sequences in the speakers' minds is dynamic; Wray (2002:101), for instance, notes that it changes with the needs of the speaker.

The importance of phraseology for fluent language production has been repeatedly highlighted by researchers in the field, both in ESL and EFL contexts, and most studies have advocated that learning materials give more space to prefabricated sequences. However, the sociolinguistic aspect of phraseology for learners of English as a second or foreign language is not as consistently taken into account. If formulaic sequences have currency within a speech community (Wray, 2002), we may add that, as a consequence, using native-like formulaic sequences entails being part of the speech community in question and sharing the same culture and values.

Most corpus-based and experimental studies actually originated in English-speaking countries, and some degree of sharing is therefore implicit in the physical presence of the learner in the community. However, the same cannot be maintained for EFL contexts, especially in the case of the early learner corpora. For example, Gilquin and Paquot (2007) observed that the EFL learners who contributed to the International Corpus of Learner English (ICLE) in the 1990s had not had much exposure to authentic spoken English outside of classroom. It is also worth pointing out that recent studies indicate that, in terms of pragmatic competence, learners show considerable progress only after about nine months or more of active participation in the speech community (Schauer, 2009). With the exception of Wiktorsson (2001), who mentions the influence of English language television and movies on her Swedish learners, most accounts of learner language through studies of written or spoken corpora carried out in EFL contexts, overlook the social aspect of phraseology, perhaps considering it implicit.

The reason for this reticence may well reside in the 'paradox' of native-like fluency. The acquisition of a second or foreign language with native-like characteristics is widely judged to be unattainable. Incidentally, the use of recurrent phrases plays a key role in this assumption. According to Mair (2007:456), advanced learner English is mostly grammatically correct and lexically rich but hardly ever fully natural or idiomatic. Among the reasons for this is the fact that collocational preferences reflect 'a community's attitudes and pre-occupations' much more than isolated sounds, words or constructions. According to Mair, 'fully natural and idiomatic use of collocations cannot be expected in any learned variety, regardless of the degree of institutionalisation of English in the community.' (*Ibid.*:450).

The 1989 analysis of formulae used by speakers of Singapore English carried out by Kuiper and Lin (1989) provides revealing insights into the formulaicity-culture interface. Kuiper and Lin find that the Singapore English formulae are often translations of Cantonese equivalents and that the cultural information they encode is so context-bound that assumptions based on standard English would not apply to the Singapore variety, possibly resulting in misunderstanding.

The cultural load of phraseology, however, is not restricted to formulaic expressions, and it has also been shown in collocations. Studies concerning semantic and cultural approaches to phraseology such as the articles collected in Skandera (2007) and the analysis of cultural keywords by Wierzbicka (2009) indicate that 'semantic study of English collocations can tell us a great deal about English speakers' culturally shaped ways of thinking' (*Ibid.*:102). In other words, collocations are 'a whole cloud of culture condensed in a drop of phraseology' (Wierzbicka, 2007b:50). Along similar lines, Stubbs (2009) argues that sequences usually express taken-for-granted cultural meanings; and the same belief has been expressed in the context of cognitive linguistic theory by studies such as Sharifian and Palmer (2007).

The studies collected in Skandera (2007) provide novel insights into the con-

nections between formulaic sequences and culture. Investigations of collocational preferences of familiar words reveal culturally specific preoccupations and values (see Wierzbicka, 2007a; Peeters, 2007; Bednarek and Bublitz, 2007; and Schönefeld, 2007) but can also be valuable keys for distinguishing varieties of English (see Wolf and Polzenhagen, 2007; and Mair, 2007). Examining the phraseology of tourism, Gerbig and Shek (2007:319) point out that 'frequently used linguistic routines in a particular area of meaning are as inseparably linked to the cognitive schemata the language users have formed about something, as to institutionalised cultural facts'. For example, Bednarek and Bublitz's analysis of recurrent sequences employing the verb *enjoy* shows that stable expressions, fixed syntax and invariable semantic content bring with them a 'gradual entrenchment of the cultural pattern of *having fun* as a natural and fundamental socio-cultural asset in US and UK cultures.' (Bednarek and Bublitz, 2007:129). As a consequence, formulaic sequences can be a means of implementing cultural concepts, which is exemplified by their exploitation in persuasive texts from media communication, advertising and marketing.

## 2.4 Recurrent Sequences in Spoken and Written English

According to Hyland (2008b), extensive use of pre-fabricated sequences 'helps to signal the text register to readers and reduce processing time by using familiar patterns to link elements of new information'. As corpus findings suggest that sequences are used for text or discourse management, research has focussed on finding out which sequences are preferred in different registers. The following sections give an account of studies of recurrent sequences in native-speaker writing and speech, focussing on corpus-derived findings. In EFL contexts, native-speaker corpora have been analysed mainly for the purpose of comparing their features to those of learner English corpora. This last strand of the research is dealt with,

in greater detail, in the section on recurrent sequences in non-natives (2.5.2).

### 2.4.1 Recurrent Sequences in Academic Writing

A high percentage of the studies regarding recurrent sequences in writing deals with academic text-types, such as published research articles, PhD theses and student assignments. In general, research studies have investigated frequency, structures, distribution and variation by discipline. A number of studies of sequences in native-speaker conversation have also been carried out in academic contexts. Biber and colleagues, for example, have investigated university teaching and on-campus exchanges (Biber *et al.*, 2004; Biber, 2006; Biber and Barbieri, 2007).

Research has demonstrated that academic writing is highly repetitive in terms of sequences. Recurrent sequences are used to frame discourse, they are generally non-idiomatic, and often structurally incomplete, therefore they may not be perceptually salient. In other words, recurrent sequences are unnoticed by readers and probably by writers themselves too. Findings also indicate that different registers and scientific subjects typically employ different sequences and that texts produced by professional academics do not use the same sequences as those written by novice writers and undergraduate students.

Studies of lexical bundles conducted by Biber and his colleagues have investigated different types of published and novice academic writing and speech in a variety of academic contexts (see Biber *et al.*, 1999, 2004; Cortes, 2004; Biber, 2006; Biber and Barbieri, 2007; Biber, 2009). The vast majority of the lexical bundles identified using corpus-driven methods are used as pragmatic heads for phrases, framing parts of the discourse and introducing new information (Biber, 2009). Far from being idiosyncratic repetitions by specific writers, they have systematic referential functions; for example, they serve as the building blocks of stance and reference. Biber *et al.* (2004) consider that the high frequencies of lexical bundles extracted from corpora of academic writing and the fact that they

are consistently functional are sufficient proof of their being prefabricated, and Biber and Barbieri (2007) claim that they cannot be considered mere products of statistical chance.

Bundles are ubiquitous in academic genres: Hyland (2008a) identified 240 different 4-word lexical bundles in research articles, PhD theses and Masters' dissertations across four different disciplines. He found considerable variation in disciplinary preferences, with greatest similarities between Business Studies and Applied linguistics. Apprentice writers do not use the same lexical bundles as professional authors; and when they do, they use them differently. Using a similar methodology, Cortes (2004) compared lexical sequences used in published academic works in history and biology with the sequences used by students in the same disciplines and found that there is a functional mismatch between the two types of writers. Some bundles used by students deviate from the concrete style that is typical of the published academic writing analysed by Cortes.

According to Hyland (2008b), there are psycholinguistic advantages of using repetitive frames; the widespread use of prefabricated sequences in academic written genres, in fact, contributes to the readability of a text. Hyland groups bundles in academic writing in three broad discourse functions: research-oriented, text-oriented, and participant-oriented bundles. Science writing involves actively participating in a discourse community and creating one's 'discoursal self' (Laane, 2011). Research suggests that rhetorical functions are carried out precisely through the use of lexical bundles.

Findings by both Cortes (2004) and Hyland (2008a) indicate that professionals have a wider range of sequences, while novice writers' use of sequences relies on the repetition of a restricted number of bundles. The negative effect of repetition is that apprentice writers' use of bundles makes the discourse sound redundant (Cortes, 2004). Cortes suggests that students may be trying out the language they are in the course of acquiring. Alternatively, this tendency might be explained by students clinging on to what have been labelled their 'islands of reliability'

(Conklin and Schmitt, 2008), that is, expressions they know and feel confident using. Interestingly, this particular feature is shared by novice writers and English language learners (see De Cock, 2004; and Granger, 1998). De Cock and Granger, however, warn that evaluations of repetitiveness should always consider the specific lengths and purposes of the different academic genres, professional and pedagogical. Published research articles and master's level dissertations, in fact, do not have the same lengths and purposes, which might explain differences in repetitiveness of bundles.

From an acquisitional perspective, it is also worth noting that Cortes (2004) found that student writers' bundles are used with meanings and in contexts that are more typical of speech than of writing. Once more, this turns out to be consistent with findings on learner writing, which seems to be characterised by features more typical of speech (De Cock, 2004). Cortes (*Ibid.*:415), however, states that in her study the use of academic bundles by students did not show consistent development across levels, even though some faculties provide intensive writing courses in the disciplines. However, these speculations are based on the comparisons of writings by different students at different levels and a proper longitudinal study was not carried out.

The influence of classroom teaching and textbooks as another possible explanation of novice writers' use of lexical bundles is hinted at in Biber *et al.* (2004) and in Biber and Barbieri (2007). Biber and colleagues found that 'classroom teaching incorporates both 'oral' lexical bundles and 'literate' lexical bundles' (Biber *et al.* 2004:379), meaning that professors tend to mix the more informal spoken register with expressions from academic writing. This mixing of registers might have an influence on undergraduate students' writing. A similar argument is put forward for textbooks, a source of much of the students' knowledge of the discipline, because even though their purpose and production is similar to academic prose, content is presented in a way that is accessible and engaging to students, using language that is more conversational than academic.

Hyland (2008b:5) suggests that prefabricated expressions denote fluent production because they 'are familiar to writers and readers who regularly participate in a particular discourse'; so a competent participant in a community uses bundles in a 'natural' way, while a novice does not. It might be argued that this sensitivity to the preferences of competent writers will eventually be acquired by those novices who wish to become part of that speech community, which may not be the case for all the undergraduates in Cortes's study. For example, recent studies of adolescent writing in the US, such as Kibler (2011), report that students 'may not see themselves in the writing roles that teachers envision for them' (*Ibid.*:211) and that this is strictly connected to wider social and linguistic issues.

In order to test the hypothesis that idioms are acquired throughout one's lifetime, Minugh (2008) analysed the COLL corpus of online newspapers written by college students from different English-speaking countries. He looked for specific idioms from the *Collins Cobuild Dictionary of Idioms* (2002) and, as hypothesised, he found a number of differences in the use of idioms both geographically, between text types and in the types of idioms young online journalists seem to prefer. Their marked preference for idioms that come from proper names, 70% of which refer to the entertainment industry (i.e. films, songs, TV shows, computers, and so on), clearly indicates the cultural information needed to be competent participants in the 'student community' they are writing for. Idioms, like prefabricated sequences and lexical bundles, therefore, are a component of the sense of belonging, or wishing to belong, to a specific speech community.

From a cultural perspective, Schmied (2011a) seeks to unify academic writing and New Englishes bringing together language, cognition and culture. Awareness of these three dimensions is important to users of English in these contexts, as they have 'no native-speaker intuition and [...] may consciously or subconsciously deviate from the codified (native) norms' (*Ibid.*:9). Schmied argues that, like New Englishes, academic written language is subjected to both centrifugal and centripetal forces, alternatively supporting national tendencies

and uniformity of style. Perhaps, no other professional lingua franca is so globally widespread and, at the same time, so tightly controlled by 'gatekeepers' such as journal editors and reviewers. Schmied concludes by proposing the implementation of a module on 'English as an international academic language and academic writing' for European and international students.

### 2.4.2 Recurrent Sequences in Native-speaker Speech

Biber *et al.* (1999) is a significant work also because it compares lexical bundles across different registers, focussing specifically on their use in the two registers that are the furthest apart: conversation and academic writing. Overall, conversation is characterised by a larger stock of lexical bundles compared to academic writing. It also features repeated local expressions (Biber *et al.*, 1999:998), which explains the fact that it is often considered overly repetitive and formulaic. Its repetitive repertoire, as reported in Section 2.2, is explained in terms of real-time production constraints, which are said to encourage the use of prefabricated sequences. In general, due to its interactive nature, conversation is rich in expressions of politeness, emotion and attitude and these functions are evident in the types of lexical bundles that occur most frequently.

The *LGSWE* (1999) groups lexical bundles into 14 different categories according to their grammatical structure. Conversation bundles mainly revolve around personal pronouns, active verb phrases, question fragments, wh-clauses, to-clauses, that-clauses, adverbials, prepositional phrases, noun phrases and quantifiers (*Ibid.*:1001-2). The most frequent pronoun is the first person pronoun, usually accompanied by a stative verb, while *you* is often followed by the main verb *want* and is normally part of interrogative or conditional clauses. In terms of functions, personal pronoun bundles function as 'utterance launchers' and present personal stance. A large number of these bundles include negative verb forms and state verbs. Discourse markers *I mean* and *you know* also figure among the most frequent bundles of conversation.

The 'extended verb phrase with active verb' category is particularly interesting in terms of formulaicity and includes expressions like *have a cup of tea*, *hang on a minute*, *get on with it*, *haven't got a clue*, etc. Question fragments are characterised by the presence of verbs which are among the most frequent in the language (such as *have*, *like*, *want*, *tell*, *know*, etc.) and are usually included in questions asking about the needs or desires of the addressee. Among the most frequent adverbial clause bundles are the formulaic sequences *as far as I*, *as soon as you* and *as long as you* (*Ibid.*:1011).

Hedging and vague reference are typical functions of noun phrase bundles, with expressions like *and things like that*, *and that sort of thing* and *nothing to do with*. Several prepositional phrase bundles are place or time adverbials, while quantifier expressions are generally fixed and have emphatic functions, such as *all over the place*, *and all the rest of it* and *all of a sudden*. Pure idiomatic expressions, however, are not frequent and Biber *et al.* also note that conversation is characterised by the repetition of meaningless sound bundles, like *mm*, which are used to mark agreement or affirmation (*Ibid.*:1014).

Another spoken genre that has attracted the attention of research on recurrent sequences is academic speech. Academic speech includes all the spoken interactions taking place in a university context and academic lectures are but one example. Academic lectures have been investigated by Nesi and Basturkmen (2006) and Biber (2006), while Biber and Barbieri (2007) used the T2K-SWAL Corpus to look for lexical bundles in other spoken university registers, such as classroom management, meetings of individual students and faculty members, study groups and on-campus service encounters.

Classroom teaching and traditional academic lectures use a large set of different lexical bundles (Nesi and Basturkmen, 2006). They carry out important discourse organisation functions and are used to indicate logical relationships. In academic lectures, for example, they are used as signals of transition between topics. Biber *et al.* (2004) also found that lectures mix conversational and formal bundles,

sometimes preferring the conversational alternative, perhaps to be more direct or make the discourse more lively (see, for example, *and this is* for *namely*, *at the end of the day* for *finally* and *and you can see* for *in other words*). It is thought that this oral style might have an influence on the language produced by students in written texts.

Biber's study identifies three primary discourse functions: stance expressions, discourse organisers and referential expressions. In addition, he develops a taxonomy of sub-functions, working inductively from analyses of concordance lists. Results indicate that in spoken university registers stance bundles predominate, however, each university register has its own specificity. For instance, in study groups epistemic bundles (e.g. *I don't know what*, *I don't know how* and *I don't think that*) are prevalent. On the other hand, desire bundles (e.g. *if you want to* and *I don't want to*) are frequently found in classroom management, and intention bundles (e.g. *we're going to* and *I'm not going to*) are present only in classroom teaching (Biber *et al.*, 2004).

Simpson and Mendis (2003) analysed MICASE (Michigan Corpus of Academic Spoken English), a corpus of contemporary speech recorded at the University of Michigan between 1997 and 2001, containing 1.7 million words of recorded speech. Looking for evidence of the use of idioms in spoken academic contexts, Simpson and Mendis found 238 idiom types, 123 of which occurred only once. In particular, their results revealed that idioms are used idiosyncratically by some speakers, who repeat them often, while other speakers do not use them at all (*Ibid.*:437).

Simpson and Mendis also carried out a qualitative analysis of the idioms in order to identify their main functions. In academic speech, idioms are used for evaluation, description, paraphrase, emphasis, collaboration and metalanguage. Particularly informal and slang idioms were often employed to reduce the formality of academic discourse or to create a sense of solidarity and collaboration. The idiomatic metalanguage used in the corpus, instead, had text-organizing

functions, with idioms functioning as signposts that signal logical connections. It should be noted that these results are consistent with findings by Biber and colleagues and seem to indicate that discourse organization is closely connected with repetitive sequences of words.

## 2.5 Recurrent Sequences and Second Language Acquisition

The interest in the phraseological quality of written and spoken English produced by native-speakers described in the previous section has fuelled studies of recurrent lexical sequences in the language of non-native speakers. The following sections review relevant research carried out in different contexts and using different methodologies.

The research can be divided according to the methodologies used (studies of recurrent sequences based on learner corpora and experimental studies), the learning contexts analysed (research originating in countries where English is learnt as a second or as a foreign language) and their focus on different registers (studies dealing with written or spoken learner production). Empirical studies mainly employ comparisons of recurrent sequences in native and non-native language production, with some examples of comparisons between different learning populations and, increasingly, across more or less institutionalised varieties of English.

The general aims of this research field are to shed light on the holistic storage of sequences in the mental lexicon, to understand second language development and identify the different factors influencing second language production. In terms of findings, overall, studies have demonstrated learners' problems with collocation, the tendency to repeatedly employ a restricted number of sequences and to mix the spoken and written registers. In addition, some studies have been aimed at developing pedagogical applications and teaching materials suitable for the

various contexts where English is learnt.

Given the increasing availability of written and spoken learner English corpora collected by different institutions and made available to the international research community, and the development of sophisticated software used to analyse the data, over the past twenty-five years corpus methodology has been the preferred choice for research into learner phraseology. For practical and methodological reasons, the following review considers studies which have an immediate importance for the present research, with a specific focus on the publications that have originated from the ICLE project.

### 2.5.1 Recurrent Sequences in ESL and EFL

Early studies of phraseology, such as those reviewed in Section 2.1, were groundbreaking because they stressed the importance of sequences for L2 proficiency and the fact that these sequences needed to find an appropriate space in the English language classroom. However, they did not study learner language empirically and systematically. The necessary next step, therefore, was quantitative assessment of phraseology in learner English.

Empirical studies of acquisition and production of formulaic sequences and routines have been carried out in both ESL and EFL contexts. The main difference between ESL and EFL contexts is that in the former context, data collection takes place at a moment when learners and native-speakers are part of the same speech community, while in the latter, the English language has no currency in the speech community and is mainly encountered in institutional settings as the object of instruction.

Bridging the gap between EFL and ESL, Cobb's (2003) replication of European research in a Canadian context confirmed EFL findings, indicating that the way English language learners acquire lexical sequences is essentially the same in both contexts. In fact, while in the past ESL and EFL were seen as totally separate, the distinction is nowadays less clear-cut, and the boundaries are in the process of

being redefined. Traditional boundaries between ESL and EFL have become more blurred because studies have started to include L2 data collected in institutional settings in English-speaking countries (i.e. American and British universities) and L2 data by learners who live in countries where English is institutionalised as the medium of instruction and used intranationally for a variety of functions (i.e. communication with speakers of different L1s, language of the media, and so on), such as the Tswana variety of South Africa included in ICLEv2.

Gilquin and Granger addressed the ESL vs EFL dichotomy in Mukherjee and Hundt (2011) and, using supporting evidence from ICLE, highlighted the importance of multiple factors in learners' productions.

> 'Although ICLE is essentially an EFL corpus, it is important to bear in mind that there are a number of factors that blur the line between the two situations, amongst them the presence or absence of language instruction (in the case of ESL), the number of years of instruction, the focus of language lessons (focus on form and/or communication), the use of the target language for some or all of the non-language subjects (for EFL), the quality of teacher talk, the type and amount of exposure to the target language outside the classroom, in particular access to English-speaking media and in the case of EFL learners, the amount of time spent in a country where English is spoken.' (*Ibid.*:57)

Due to different levels of exposure to authentic English in different settings and through different media, today ESL and EFL are increasingly considered on a continuum with in-between categories. For example, in a study of collocations with the preposition *into* in ICLEv2, Gilquin and Granger (2011) found that, by setting the time spent in English-speaking countries to less than three months, the high levels of exposure of Dutch learners to television and films in English placed them at the end of the cline closest to native (British) speaker norm. Spanish learners, on the other hand, could be positioned at the opposite end of

the same cline, as their exposure to authentic English is inferior, due to the fact that they live in a 'dubbing country'; where films and television programmes are regularly dubbed into the Spanish language.

ICLEv2 includes language produced by Tswana learners of English. Findings indicate that their English displays features of both EFL and ESL. Their position along the cline could be considered on a par with that of speakers of other L2 varieties of English from countries in the outer or norm-developing circle described by Kachru (1985) because, depending on the feature selected for study, it is sometimes the closest to and sometimes the most dissimilar from native English.

Reflecting on Crystal's (2003) observation that collocations are likely to be one of the features differentiating varieties of English, Nesselhauf (2009) set out to investigate co-selection phenomena in different varieties of English on the basis of the ICE-corpora of Kenyan, Indian, Singaporean, and Jamaican English. As she noticed that some collocational phenomena described for some institutionalized L2 varieties of English resembled those that have been described for learner English, her study included learner texts from ICLE. Some similarities were found, especially when co-selection phenomena concerned language-internal irregularities in the phraseology of standard English. In other words, when it comes to collocation, both ESL speakers and learners of English were found to produce language using similar strategies. In particular, both have a tendency to regularise language irregularities. For example, 'when verbs behave irregularly because they do not take a preposition while closely related words and expressions do, the preposition found in related expressions is added to the verb through different means' (*Ibid.*:23) (as exemplified by the collocations *request for* from *ask for*, or *discuss about* from *speak/talk about*). L2 speakers also tend to economise on learning effort by reducing the meanings, contexts of use or variability of collocations (see *play a role* vs *play a part*), or to explicitly express the direction in verbs of movement, even when it is already present in the meaning of the verb (see *enter into* and *return back*). Interestingly, through spot checks in the

British National Corpus, which is commonly considered one of the yardsticks used to measure nativeness, Nesselhauf found some of the 'new' prepositional verbs (*Ibid.*:21). These strategies can be traced back to L1 transfer only in part, and Nesselhauf concludes that 'co-selection phenomena that only display a low degree of idiomaticity and culture-boundedness tend to have similar characteristics across L2 and learner varieties' (*Ibid.*:22-3).

Nesselhauf's new strand of research seems to be spreading in learner corpus linguistics, another such example being Götz and Schilk (2011). Based on the assumptions of the increasing fluidity of context of acquisition, epitomised by the current the redefinition of ESL and EFL contexts, on the sociolinguistic notion of speaking community, and on the universal nature of language acquisition, Götz and Schilk compared native, second language and learner varieties by analysing 3-grams in the ICE-GB, LOCNESS, ICE-India and LINDSEI-GE corpora. Their quantitative analysis revealed significant differences in the number of 3-grams in native-speakers and learners, corroborating previous evidence that learners rely on a restricted repertoire of sequences. Their qualitative analysis, instead, brought to light interesting findings regarding the high percentage of the 3-grams *I would like* and *would like to* in ICE-India and the use of the 3-gram *I don't know*. Götz and Schilk found that *I would like* and *would like to*, which normally appear in contracted forms in ICE-GB, are frequently used in their full form in ICE-India. *I don't know*, instead, is mainly used to express uncertainty about language correctness rather than with the hedging function prevalent in native British English. Even though the authors caution the reader to duly take into account differences in the design of the corpora analysed, their findings in terms of tokens and functions are particularly relevant to the present study and are discussed in Chapter Six.

All things considered, in the field of phraseology the main differences between ESL and EFL contexts seem to hold strong. In ESL varieties English is acquired through the spoken medium in early childhood and therefore with a more holistic

approach, it is the language of school instruction and of the media, and it may also be used in a variety of formal and informal settings, depending on the country. However, modern EFL contexts also exist in a variety of forms: English is increasingly being taught to young children in and out of school, it is increasingly used as the medium of instruction for other subjects (as in Content and Language Integrated Learning, or CLIL) and it is acquired through increasing exposure to English-language media (thanks to the Internet, as we shall soon see). The ESL/EFL dichotomy, therefore, may soon need rethinking and reformulation.

The latest means of exposure to authentic language for learners throughout the world is computer-mediated communication. According to Herring (1996; 2002; 2010), CMC is an important new means of communication which exhibits features of its own (see Collot and Belmore, 1996). The language of Cyberspace is currently reshaping the notion of 'speech community' and transforming it into a fluid concept able to transcend geographical boundaries, as it creates online multilingual communities using English as the language of communication. CMC and its linguistic features are key to the present study of sequences in learner English and its status as a new variety of English used for intercultural communication is discussed in Section 2.6.

For all the reasons above, the current concept of native and non-native contexts might turn out to be outdated, as the main difference seems to reside in the different statuses of the circles of English. The 'norm-providing' or 'norm-developing' status of the inner and outer circles of English, compared to the 'norm-dependent' status of the expanding circle, accounts for the fact that phrasal expressions or collocations come to be considered non-standard, creative manipulations of the language in World Englishes, while the same expressions would be judged as grammatical or lexical mistakes in EFL norm-dependent contexts and therefore be considered signs of unsuccessful language acquisition (see Götz and Schilk, 2011). Perhaps, as suggested by Gilquin and Granger (2011), we should substitute the concept of Learner English with that of Learner Englishes.

### 2.5.2 Features of Learner English Sequences

There is widespread agreement that the acquisition of recurrent sequences in first and second language follows similar developmental steps (as described in Wray and Perkins, 2000 and Wiktorsson, 2001, see Section 2.3). It is therefore hardly surprising that 'native-like use of phraseology and idioms' is placed among the indicators of advancedness for L2 learners. Even though there is as yet no universally accepted definition of advancedness, as summarised by Callies (2008), what scholars seem to agree about is that learning English well equates with learning the appropriate use of sequences (Howarth, 1998). On the one hand, formulae make communication smoother because of processing advantages, i.e. higher reading speeds for written texts, for both natives and non-natives (see Conklin and Schmitt, 2008); on the other hand, formulaic sequences are expected and accepted routines of a speech community (compare Hyland, 2008b).

From the psycholinguistic perspective, claims about formulaic language being stored and processed holistically have been investigated in non-native speakers as well as in native speakers. The validation of this oft-cited claim with empirical findings has been the object of studies by Ellis *et al.* (2008), Conklin and Schmitt (2008) and Nekrasova (2009). For learners, the advantage of using prefabricated sequences lies in saving processing effort and freeing up the working memory, which can concentrate on the construction of fluent discourse. Ellis *et al.* (2008) reported on two experiments investigating the psycholinguistic validity of the formulae for both native and non-native-speakers. Their results show that, from a processing point of view, advanced learners immersed in English studies in American universities for a period of three to twelve months are sensitive to frequencies of formulaic sequences. Similar results were reported for an experimental study on the processing of idioms included in cloze-tests by Conklin and Schmitt (2008). They determined that learners do not find formulaic sequences more difficult to understand than literal speech.

Other experimental findings concerning L2 speakers, however, appear to con-

trast this. Siyanova-Chanturia *et al.*'s (2011) study employing eye-tracking technology to investigate the on-line processing of idioms investigated L2 learners as well as native speakers. Findings suggests that L2 speakers process idioms and novel phrases at a similar speed and that figurative uses are processed more slowly than literal ones. A recent experimental study by Nekrasova (2009) is particularly relevant to the present one, because it focusses on lexical bundles, albeit in an L2 academic context. Nekrasova used a gap-filling activity and a dictation task to investigate L1 and L2 speakers' knowledge of different types of lexical bundles. Her findings suggest that for L2 learners referential bundles and discourse organising bundles have different levels of psychological validity and that salience is also influenced by register distribution. In other words, lexical bundles are recognised and produced by learners if they appear frequently in a specific register and with a specific discourse function. These three factors concur to appropriate production of sequences. In addition, it was found that lower proficiency learners did not produce as many lexical bundles as L1 speakers and higher proficiency learners.

Such experimental studies are important as they can be instrumental in confirming or rejecting assumptions formulated on the basis of corpus findings. Researchers such as Gilquin (2005) and Gilquin and Gries (2009) advocate the employment of a combination of corpora and experimental methods to validate corpus-based results against other types of findings. It could be argued, however, that these methodologies have much in common: both analyse English L2 production with a view to gaining more insights into the learners' mental lexicon and the principles underlying second language acquisition, using L2 samples or texts produced in institutional or learning settings, such as universities and ELT or EAP classrooms.

The majority of the research studies on recurrent sequences in learner English have their origins in the ICLE project, started by Granger in 1998, and in the publication of the two corpora assembled at the University of Louvain, ICLE

(2003) (followed by ICLEv2 in 2009) and LINDSEI (2010). Since Granger's study of prefabricated patterns in advanced EFL writing published in 1998, issues of 'phrasal chunkiness' in learner production were explored mainly by De Cock (1998, 2000 and 2004). The same path was followed by researchers from different European Universities using ICLE, or LINDSEI or focussing on their respective national subcorpora (see among others Wiktorsson, 2001 and Aijmer, 2009b for Swedish learners; Neff *et al.*, 2004 and Rica Peromingo, 2010 for Spanish learners; Gilquin and Paquot, 2008 for French learners; Callies, 2009 and Brand and Götz, 2011 for German learners; and Hasselgard, 2009a for Norwegian ones). More recently, new learner English corpora made their appearance in research articles and publications and, as summarised in Section 3.2, their quantity is on the rise.

Most corpus-based studies of learners' sequences involve evaluation of their written or spoken production by means of comparison with native-speakers, usually employing a native-speaker comparable corpus to quantitatively determine linguistic differences between natives and learners. As a consequence, the main findings about recurrent sequences in learner English are described in terms of 'overuse' and 'underuse'. These terms are 'descriptive, not prescriptive, [...] [and] merely refer to the fact that a linguistic form is found significantly more or less in the learner corpus than in the reference corpus.' (Gilquin *et al.*, 2007:322). These descriptive categories, together with the more prescriptive term 'misuse', are used in most of the literature on learner English.

Features of learner use of sequences identified through comparisons include: repetitive use of a restricted range of sequences, extensive use of markers of involvement, lack of register awareness, and idiosyncratic uses of phraseology. In other words, learners have preferred sequences and rely on well-rehearsed routines to build their utterances (see De Cock, 1998, 2000 and 2007), and their phraseology is therefore characterised by repetitiveness. These performance problems related to phraseology have been found to be partly due to developmental factors, partly resulting from L1 transfer, and partly teaching induced.

The next two sections (2.5.3 and 2.5.4) focus on studies of recurrent sequences in written and in spoken learner English. Due to their relevance to the present research, it reviews the findings from European corpora in more detail, even though mention is made of results deriving from American and Asian corpora.

### 2.5.3 Studies of Sequences in Learner Writing

Corpus-based studies of recurrent sequences in learner writing using corpora have brought to light the inherent phraseological quality of English learner language. This should not come as a surprise, since we know that learners memorise formulae and routines right from the beginning of their language learning. Usually, they rely on fixed formulae to carry out a variety of pragmatic functions. More importantly, though, research has revealed that, at later stages of acquisition, learners produce native-like sequences (Nesselhauf, 2005) of different types and with various functions and they employ them as building blocks of discourse (De Cock, 2004). Still, studies of learners' formulaic sequences in written texts report that learners' use of word combinations is what makes learner writing stand out as unnatural, odd, or foreign.

Due to the increasing importance of EAP, stemming from the large number of non-native students in many US and UK universities, learner corpus research has focussed mainly on argumentative essays and academic writing. The widespread use of academic corpora is also partly connected to the ease with which such corpora can be constructed in university contexts, given the availability of learner writings in the form of essays, assignments, exams, and so on.

Overall, findings on learner sequences from texts produced in academic contexts have highlighted failures in academic register awareness. In other words, learners include frequently used, native-like sequences, but they generally choose sequences that are not typical of the academic register. A high percentage of the findings in this regard comes from ICLE (2002) and several of its national subcorpora (Wiktorsson, 2001 and Rica Peromingo, 2010 *inter alia*).

Studies agree that learner English is characterised by *repetitive phrasal chunkiness* (De Cock, 2000). That is to say, learners' restricted formulaic repertoires lead them to overuse some sequences, which makes their texts sound non-native because their phraseology is less diverse than that of native-speakers. On the other hand, this overuse creates an effect of verbosity. Granger (1998), for example, illustrates this verbosity in the overuse of sequences such as *the fact that* or *as far as X is concerned*.

When Wiktorsson (2001) investigated the 10,000-word Swedish component of ICLE, consisting mainly of argumentative essays on topics such as the environment, immigration and inventions, she found that Swedish learners' prefabricated expressions indicated a more spoken style. She connected the finding to exposure to English and American television, as young people in Sweden get most of their English input through this media (*Ibid.*:10).

Similar findings were related by Gilquin and Paquot (2008), who sought specific words and phrases from an Academic Keyword List in the second version of ICLE. They found that, overall, EFL learners' argumentative writing shows a clear influence of speech, even if these learners mainly learn the language through formal instruction. Among the possible causes of these surprising results, they mention teaching materials, transfer of register features from the L1 and exposure to television and the Internet. In addition to this, however, there are developmental factors that should be taken into account. As the authors point out, the subjects are novice writers and orality in tone is typical of L1 production in novice and adolescent writers.

Gilquin and Paquot (2008) convincingly argue that professional academic writing is not strictly comparable to EFL learner writing and should not be used to evaluate interlanguage, even at advanced levels. For comparison with native-speaker production, they deem it considerably more appropriate to use ad hoc corpora of writing by novices, such as the Louvain Corpus of Native English Essays (LOCNESS), consisting of 300,000 words of argumentative essays by British

university students, grammar school students taking their A-levels, and US college students.

It might be argued, however, that argumentative essays of the kind collected in ICLE cannot be classified as an academic text-type. It should also be borne in mind that some of the texts included in ICLE were produced for language assessment purposes, or they were written exams produced under time pressure. These conditions are not comparable to academic writing, neither of the professional, nor of the novice kind. Another point worth mentioning here is that the amount of writing learners are required to produce in some university contexts, certainly in the Italian one, is scarce both for their native and for their second language. It follows that most of the EFL writers whose samples make up ICLE-ITA can be considered extremely inexperienced writers of academic texts.

In order to carry out fair comparisons between academic writing by native and non-native writers, several universities in EFL countries collected their own corpora of learner academic writing. Ädel and Erman (2012), for example, investigated the one million word SUSEC (Stockholm University Student English Corpus) a corpus of academic writing by Swedish students, and Callies and Zaytseva (2011), at Mainz University, compiled CALE (a Corpus of Academic Learner English by advanced students of different language backgrounds). CALE was created to be able to analyse written academic English belonging to different genres and to be comparable to native writer corpora such as MICUSP (the Michigan Corpus of Upper-Level Student Papers) and BAWE (the British Academic Written English corpus).

Comparisons of native and non-native academic text-types have also been carried out in ESL university contexts, both in Britain and in the United States. In general, they have revealed a gap between expert academic prose, L1 student academic writing and L2 student academic writing. At Lancaster University, Baker and Chen (2010) compared three small corpora of academic writing to identify similarities and differences in recurrent word combinations at different levels of

writing proficiency. One corpus contained writing from L1 Chinese learners of English. The most marked differences were the use of referential expressions, especially stance bundles, which were not used by the Chinese learners. Findings indicated that learners also underuse epistemic bundles and hedging devices, showing a 'tendency to be categorical and to over-generalize' (*Ibid.*:43). This is consistent with observations by Allen (2009) about the 'more authoritative tone' of Japanese writers, and it was found to be true of learners from a variety of L1 backgrounds. Baker and Chen (2010) also found a relation between recurrent word combinations and writing proficiency. They found that 'the number of recurrent word combinations increases with advancing writing proficiency, which is the case both for the range of lexical bundles used (types), and the overall occurrence of lexical bundles (tokens).' (Baker and Chen, 2010:43). However, issues of comparability are extremely important in this respect, as frequencies of recurrent word combinations are strictly connected to corpus size and results from different corpora may not be fully comparable if the corpora were not designed to be comparable in the first place.

Differences are also reported for writer visibility and subjectivity. Petch-Tyson (1998) identified this feature in personal pronoun references, while Milton (1999) found that Chinese L2 students used the formulaic sequence *in my opinion* more frequently than UK students. Findings consistent with this were reported by Cobb (2003), who noted that learner texts display considerable personal involvement, and McCrostie (2008), who found high percentages of the sequence *I think* in Japanese learner writing. Neff *et al.* (2004), on the other hand, show that excessive writer visibility is also typical of native writing. The analysis of ICLE by Gilquin and Paquot (2007) showed that

> '...learners overuse a number of expressions which make them particularly visible as writers, cf. *I think*, *to my mind*, *from my point of view* and *it seems to me* to express a personal opinion, and *I would like to/want/am going to talk about* to introduce a topic or an idea.'

(*Ibid.*:4)

Clearly, the learner writers show that they do not possess a 'confident and expert mind in full control of the material' (McCrostie, 2008:110), in other words, they do not feel part of the academic community whose discourse they should emulate. Similarly, Allen (2009) explains learners' use of *it is known* as a preference for the presentation of proven facts rather than arguing with theories, which is typical of inexperienced writers.

Applying results of research to pedagogy and therefore providing learners with enhanced learning opportunities is widely considered critical (see Flowerdew, 2001), so researchers studying recurrent sequences recommend a focus on sequences in teaching practice and materials. Studies also report that in writing the advantage for learners of using prefabricated sequences is connected to increased grammatical accuracy. Allen (2009), for instance, found that bundles in the academic writings collected in ALESS (a corpus of research papers written by Japanese students attending the English for Science Students Course at the University of Tokyo) were mainly free of errors and exhibited considerable convergence with published and native-speaker writing. Even though he warns that the Japanese learners' texts in his corpus were extensively edited and peer-reviewed, Allen considers that focusing on appropriate use of lexical bundles in texts is essential if learning is to take place. Following the same conviction, Li and Schmitt (2009) stress the positive effect of reading in their longitudinal study of a Chinese MA student over the course of ten months of study at a British university. The learner in question was reported to have learned 166 new lexical phrases, drawing on explicit and implicit sources, and to have improved in her degree of appropriateness. McCrostie (2008) also reports 'a dramatic reduction in this degree of writer/reader visibility' after the first year of study for his Japanese learners (*Ibid.*:112).

Among others, Granger *et al.* (2002a) and Gilquin *et al.* (2007) repeatedly stress the importance of using findings from learner corpora for materials design.

Right from the beginning of learner corpus analysis, Flowerdew (2001) advocated using findings on collocational patterning, pragmatic appropriacy and discourse features from native and learner corpora in the preparation of EAP materials. Access to corpus-informed materials is clearly advantageous for learners in EFL environments, which are said to be 'impoverished learning environments' in terms of lexical density and lexical sophistication of discourse, partly due to the shortcomings of non-native English teachers (Gilquin and Paquot, 2007:6).

This is a rather unflattering view of the teaching profession in Europe, and other scholars, such as Römer (2007), claim that textbooks should share the blame. It is a fact that, especially in the early years of language learning, a considerable part of learners' input comes from textbooks. Römer (2005 and 2007, *inter alia)* has repeatedly shown textbooks to be an unreliable source of authentic English discourse, both in the spoken and in the written genres. The fact that learner output shows a mix of features from speech and writing, therefore, can be explained by learner confusion stemming from 'conflicting values and influences: EFL textbook/grammar influence, real English influence, and the influence of their first language (here German), which may contain potentially concurring structures' (Römer, 2007:360).

Currently, there seems to be a gradual seeping of research into published materials, as can be seen, for example, in the corpus-informed academic writing section in the second edition of the Macmillan English Dictionary for Advanced Learners, mentioned in Gilquin *et al.* (2007). At the same time, however, the notion of nativeness in academic contexts is the object of an ongoing critical reassessment. Examples of this criticism can be found in Römer (2009) and in Schmied and Haase (2011). Römer (2009) restated the suggestion that, in order to take into account the developmental aspect of academic phraseology, the notion of expertise should substitute that of nativeness. From an intercultural perspective, Schmied (2011b) highlights the cultural dimension that distinguishes native and non-native academic writing, as linguistic choices enable English language

users to express their own culture as well as to participate in their international 'community of practice'.

To sum up this section, learners of English in non-native learning contexts are found to be weak academic writers. The reasons for their weakness have been pinned down with precision by a wealth of research studies and publications. In terms of phraseology, their use of appropriate sequences is deficient and even if they fare better in ESL contexts, they still show a fundamental lack of register awareness and the influence of their L1 remains, both in phrasal choices and in discourse culture reflected in rhetorical devices. The features of learner academic English described above also result in learner academic writing often being described as 'speech written down'. This observation is important for the discussion of CMC in Section 2.6, which has also been considered by research as 'speech written down', albeit on computer rather than on paper.

### 2.5.4 Studies of Sequences in Learner Speech

Native-like combinations are deemed necessary for fluent speech, and the more advanced the proficiency of learners, the more they are expected to be able to use phraseology in a natural, native-like way in speech as well as in writing. Several of the observations and conclusions outlined in the previous section also apply to recurrent sequences in learners' spoken production. However, the number of research studies revolving around spoken learner language is decidedly lower, due to the scarce availability of relevant corpora. In general, research studies report that learners rely on a limited number of prefabricated sequences (see De Cock, 2007 and Götz and Schilk, 2011), which tend to be repeated and sometimes used with different meanings and functions compared to native-speakers (see De Cock, 2004 and Aijmer, 2009b). In particular, learner speech is lacking in expressions of vagueness and differs in terms of hedges, revealing, once more, a mix of spoken features and more formal ones, borrowed from writing. Researchers have also focussed on learners' dysfluencies, and their use of pauses and hesitations has

been interpreted as evidence of non-holistic storage of recurrent sequences of English in the learners' mental lexicon (see Gilquin and Cock, 2011).

Foster (2001) carried out a study investigating native and non-native speakers use of ready-made sequences in a 20,000-word corpus constructed from an interactive task. She noted that 'the non-native-speakers were constructing a great proportion of their language from rules rather than from lexicalised routines, with or without the benefit of planning time.' (*Ibid.*:90). She concluded that this should be hardly surprising in contexts where classroom teaching revolves around grammar and rules. Foster found that the non-natives tended to repeatedly employ a smaller number of sequences compared to native-speakers.

Findings of this kind need to be validated by large corpus-based studies before generalisations can be considered more than teachers' intuitions. Corpora of spoken language, however, are difficult to construct and until the availability of LINDSEI, there were very few studies of learners' sequences in speech. LINDSEI and its control corpus LOCNEC (Louvain Corpus of Native English Conversation) contain informal interviews of about fifteen minutes each (corresponding to approximately 2,000 words of interviewee speech) on topics such as university life, hobbies, foreign travel, plans for the future, etc.; followed by a story telling activity based on pictures.

Among the studies based on LINDSEI, De Cock (2004) investigated the corpus in order to assess learners' reliance on 'individual bricks' rather than prefabricated sequences for building their speech. Her corpus-driven analysis compared the oral performance of fifty learners from the French component of LINDSEI with the native-speaker speech contained in LOCNEC. Her corpus-based results present a 'complex picture of underuse, overuse and misuse of target language sequences' (*Ibid.*:143); and the quantitative analysis of the two corpora found that, similarly to learners' writing, learners' speech was characterised by repetitiveness of relatively few recurrent sequences. Moreover, a great number of sequences contained repeats and hesitation items which indicated learners' encoding problems.

De Cock (2004) also carried out a qualitative analysis, which revealed critical differences between learners and native-speakers in terms of the functions of learners' preferred sequences. The study found that learners' recurrent sequences in speech were less interactional and less involved than those employed by natives and that markers of vagueness were underused by learners.

As we saw in Section 2.5.3, learners' written production tends to be informal and involved, they tend to overstate more and hedge less than native speakers (see Baker and Chen, 2010). In the case of learners' speech, the picture is more complex. For example, De Cock's (2004) study of vagueness tags revealed that learners' recurrent sequences exhibit features of formal talk. This was demonstrated by the frequent use of *and so on* and *etcetera* and by the lack of sequences such as *sort of thing*, *all that kind of thing*, *and things*, or *anything*. After highlighting the different use of the vagueness tags *sort of* and *kind of* in learners and natives, and learners' inappropriate use of *(yes) of course*, De Cock concludes that learners are 'lacking in routinized ways of interacting and building rapport with their interlocutors and of toning down and weaving the right amount of imprecision and vagueness...' (*Ibid.*:243).

Even though these findings confirm that learner English sequences lack some of the features of native-speaker conversation, the main aim of De Cock's article was to demonstrate the phraseological quality of learner speech. In her conclusion, she advocates the need of more in-depth analyses of learner spoken corpora in order to uncover learners' preferred sequences and assess the influence of L1 transfer. The formality of learners' productions in LINDSEI could also be ascribed to the task itself: i.e. a recorded interview with a time limit. As the task cannot be considered a naturally occurring piece of conversation, how the learners really viewed the task is open to speculation. Given the university setting and the teacher-led task, learners might, perhaps unconsciously, have assigned to the task a higher level of formality than wished for by corpus collectors. Clearly, when dealing with spoken production, there are so many variables at play that

generalisations should always be made with caution.

In a study of pragmatic markers in spoken English by Swedish learners based on the Swedish component of LINDSEI, Aijmer (2004:174) states that 'learners may overuse or underuse certain devices in comparison with native-speakers and therefore sound non-native'. The Swedish learners investigated by Aijmer used markers that are typical of informal conversation, such as *you know*, *I think*, *and all that*. However, these markers were found to co-occur with pauses, signalling processing problems on the part of the learner. In a more recent study, Aijmer (2009b) investigated the use of the sequences *I don't know* and *I dunno* in LINDSEI and LOCNEC, and her results show that learners who used these markers did not take full advantage of their various functions. Above all, learners employed the expressions to avoid answering questions directly or to cope with speech management problems. Aijmer (2011) reached similar conclusions in a study of learners' use of *well*, and comparable findings were reported by Müller (2005) regarding non-native use of discourse markers in general. Although, strictly speaking, some discourse markers are not sequences of words, they can be used to assess interlanguage pragmatics, which in turn has close links with formulaicity (see Granger, 1998 and Aijmer, 2004).

If repetitive use of sequences lightens processing effort and reduces processing problems, it follows that use of sequences should be a reliable measure of learners' fluency. Brand and Götz (2011) carried out a quantitative analysis of the oral production of fifty German learners recorded in LINDSEI in order to look for a possible correlation between accuracy and fluency in advanced learners. The production of five learners was analysed quantitatively, i.e. measured in terms of words per minute, and qualitatively, but neither analysis brought to light a correlation between the variables. Learner spoken production was found to be generally accurate, and mistakes mostly regarded use of tenses, aspect, and verb agreement. Greater variation between learners was found in terms of words per minute, filled and unfilled pauses and hesitations. An area that was found to

be problematic is the use of lexical phrases, a category including multi-word expressions and idioms. Despite the high error rate observed in the use of lexical phrases by German learners, no correlation was found between this measure of accuracy and measures of fluency. Brand and Götz (2011:272) concluded that 'the perception of overall oral proficiency is not based on good performance in one single variable only, but rather results from good performance across several variables'. Whether there is a correlation between use of lexical phrases and L2 fluency, therefore, remains open to investigation.

Most of the studies reviewed so far were based on the LINDSEI corpus of learner speech. An exception is Wei's analysis of COLSEC, the Chinese Learner Spoken English Corpus (in Jucker *et al.*, 2009). COLSEC, a 700,000-word corpus of oral 'episodes' from the College English Test (CET), was analysed in order to characterise the phraseological features of Chinese learners' speech. During the CET, students are interviewed by an examiner and answer questions about academic study, campus life and other topics. The test also includes group discussion sessions and is video recorded for grading purposes. After creating the corpus, Wei automatically extracted 3- to 6-word chunks and adopted Altenberg's (1998) framework to divide them into structural types. He reported frequencies that were similar to those identified by Altenberg, with chunks including clause constituents being the most frequent structural type, followed by incomplete phrases and independent clauses. Multiple clause constituents such as: *so I think*, *and I think*, *but I think*, etc., were very frequent and principally employed as sentence frames or stems. He also found that semi-fixed formulae, normally used by native-speakers to realise pragmatic functions of discourse, were greatly underrepresented. It might be argued, however, that an exam setting might not be ideal for non-native-speakers to produce these types of functions. In his conclusion, Wei admits the limits intrinsic in the nature of COLSEC and cautions against making hasty generalisations.

In general, Wei's observations are similar to De Cock's (2007:278): Chinese

learners rely on a limited number of 'thematic springboards' and their 'weakness in manipulating chunks may partly account for their non-nativeness and dysfluency'. When concentrating on a single language background, cultural influences become more apparent. Wei (2009:292), in fact, noticed the impact of socio-cultural factors in the Chinese learners' chunks he analysed, which contributed to their 'uniquely idiosyncratic formal and functional features'.

Figures 2.5.3 and 2.3 summarise the findings on spoken learner English recurrent sequences. Learner language is rich in recurrent sequences, regardless of the L1 background. Sequences are found to be mainly used to build discourse and as a strategy to gain processing time. When compared with native-speaker sequences, however, limited range, repetitiveness and misuse become apparent and learners lack appropriate sequences for pragmatic functions of discourse. The explanations generally provided by the literature include teaching practices, teaching materials, the degree of exposure to authentic spoken language, the developmental stage reached by the learners, L1 interference, motivation, and socio-cultural influences. In terms of pedagogy, language instruction methods are deemed to be crucial: where language teaching relies mostly on analytic approaches to the L2 and accuracy has more weight, the acquisition of prefabs might be reduced. It was argued, for example, that acquisition of fixed or semi-fixed formulae might be promoted in cultures (such as the Chinese one) where rote learning is still highly valued. The complex interplay of cultural, situational and developmental factors makes the study of learner language through recurrent word sequences a fascinating task for the researcher, but one that cannot possibly reward them with conclusive findings.

Today, the use of new forms of communication chiefly through the Internet and the creation of new online discourse communities add a further piece to the phraseological puzzle. The next section, 2.6, reviews studies of the features of computer-mediated communication, with special attention to studies of its linguistic features (Section 2.6.2) and of learner language produced by means of computer (Section 2.6.3).

Figure 2.2: Recurrent Sequences in Learner English Writing

1. Recurrent sequences types and tokens increase with writing proficiency;
2. Learners underuse, overuse and misuse native-speaker recurrent sequences;
3. Learners' use of sequences is repetitive;
4. Learners use sequences as building blocks of discourse;
5. Learners use frequent native-like sequences from spoken English;
6. Use of recurrent sequences increases grammatical accuracy;
7. Sequences are used appropriately if they have high frequencies and specific functions and structural patterns;
8. Learners' overuse of such sequences makes texts unnecessarily verbose;
9. Learners underuse of epistemic bundles and hedging devices;
10. Learners overuse of sequences including the first person pronoun (*I think*);
11. Learners overuse of formulaic sequences to express personal involvement (*in my opinion*);
12. Learners' use of these sequences makes them particularly visible as writers;
13. Learners' discourse organising sequences are generally appropriate to writing.

Figure 2.3: Recurrent Sequences in Learner English Speech

1. Learners rely on a limited number of prefabricated sequences;
2. Learners underuse, overuse and misuse native-speaker recurrent sequences;
3. Learners' sequences contain repeats and hesitation items which indicate encoding problems;
4. Learners use some sequences differently from native speakers (e.g. *of course*);
5. Learners' recurrent sequences are less interactional and less involved;
6. Learners lack routinised ways of building rapport;
7. Learners do not use sequences as vagueness tags and to hedge their speech;
8. Learners mix features from speech and features from writing and from formal talk;
9. Learners underuse sequences such as *I don't know* for discourse marking functions;
10. Learners overuse sequences to cope with encoding and speech management;
11. Sequences including the words *I think* are very frequent and used as sentence frames or stems.

## 2.6 Computer-mediated Communication

Since the data analysed for the present study were collected by means of a computer chat, Sections 2.6.1 and 2.6.2 are an account of the main features that characterise computer-mediated communication; while Section 2.6.3 reports on studies of learner language produced by means of CMC.

The following sections, however, are not intended as a comprehensive review of CMC studies, which would be outside the scope of the present research. Moreover, the focus is exclusively on text-based CMC, while CMC modes also include interaction by means of audio, video and static images, which may or may not be integrated with text. In text-based CMC, instead, participants interact by typing text on their computer keyboard and reading it on their computer screens. It was

the earliest type of communication among Web users and is generally carried out via email, instant messaging, synchronous chats, asynchronous discussion forums and Weblogs.

In particular, the next sections focus on discussions of the most salient features of CMC, namely: the informal, conversational quality of CMC, its distinctive language and its unique typographic features. These features are relevant to the data collected for the present study and will be discussed further in Section 4.2. Although most of the studies reviewed are about English, some mention is made of studies dealing with languages other than English, as some features of CMC have been shown to be common across languages.

### 2.6.1 Text-based CMC in English

This section is an introduction to text-based CMC in English and to its distinctive features. The linguistic impact of English CMC on CMC in the rest of the world, and the notion of the Web as an online speech community are also discussed.

As with every new technology, at the beginning, the impact of CMC on language and language change was deemed to be revolutionary. Crystal (2006), for example, reported concerns about emails being overly spoken and featuring non-standard grammar and careless spelling. Initially, additional concern was expressed over issues of linguistic homogenisation and linguistic imperialism. In general, CMC research tends to be very polarised: optimists view it as a means of global empowerment through intercultural communication, while pessimists view it as a vehicle for 'globalised sameness' at the expense of linguistic creativity, and maintain that the current dominant position of English is preventing non-English speakers from accessing the Web (as discussed in Macfadyen, 2006b and Warschauer *et al.* 2010). Many of the studies of CMC reviewed below address these initial anxieties and show that the field is currently undergoing a process of redefinition of concepts and assumptions, summarised in the works of Herring (2002; 2010; 2012), Danet and Herring (2007) and Dresner and Herring (2010).

Overall, studies of CMC have denied early worries about excessive informality, simplification of language and homogeneity. Indeed, it has been demonstrated that these features are connected to the context and function of the interaction and to the mode chosen. Murray (2000), for example, reported studies which found different levels of formality in e-mails produced by insurance company employees and by academics, with the first exhibiting a higher level of formality and linguistic complexity. Research has shown that e-mails and other asynchronous modes exhibit a great deal of linguistic variation, depending on user education and purpose of communication, while synchronous communication is characterised by the most informal style and the pressure to type at a conversational pace affects linguistic complexity. Some synchronous chats, for example, were found 'to be simpler even than spontaneous speech in terms of range of vocabulary used and measures of word and sentence length' (Herring, 2002:139).

The initial concerns about the linguistic impact of CMC were immediately complemented by a fascination with its unique features. Among the first to attract scholarly attention was the clever use of keyboard keys, which shows the high degree of adaptation of human language to new technological means. 'The myriad ways human users adapt to the constraints (and affordances) of CMC systems in order to converse' (Herring, 2010:5) include emoticons, abbreviations and non-standard spellings to imitate intonation, for humorous purposes, or to express emotional meanings. Overall, they are used to overcome CMC's lack of prosodic and nonverbal cues, even though they were found to have additional functions.

According to Herring (2002), for instance, technological constraints are not the sole reason behind CMC's distinctive features. Creativity and playfulness appear to be intrinsic to the recreational purposes of most CMC. Herring (2002:139) underlines that 'situational factors can (and regularly do) override the predispositions of the medium, and users can adapt the medium to their communicative needs, just as with communication in other media'. Visual humour, playful use

of emoticons and expressive language require more rather than fewer keystrokes, thereby flouting the oft-cited principle of economy of effort. Interestingly, linguistic and typographic creativity and playfulness have been found to be common features both cross-culturally and cross-linguistically and newcomers to an online community train themselves in their use in order to be able to blend into it (Yus, 2011).

At present, the online dominance of the English language is uncontested. Firstly it enjoys historical precedence, as the Internet originated in the United States; secondly, it has become a global lingua franca. Considerable evidence suggests that distinctive CMC features, such as abbreviations, emoticons, and conversational language, recur in non-English CMC. As Danet and Herring (2007:27) explain, some shared features 'may have originated from a single source (e.g., North American English users) and spread through contact, although some may be local responses to technical constraints of the medium'. At first, in fact, non-English CMC was hampered by the ASCII coding system, which was designed for English and made other languages difficult to reproduce. Today, technical problems have been solved for most languages, and over the past decade, the Internet has become more multilingual (Warschauer *et al.*, 2010:491) leading to the global spread of CMC.

The dominant position of English in cyberspace is also fuelled by the fact that it is the most widely studied foreign language in the world. As summarised in Crystal (2003), the status of English as the language of international communication is the result of political, economic and cultural factors which predate the advent of the Internet. Evidence is accumulating that English is the preferred choice in multilingual contexts, even when no native-speakers are present, but CMC users do not necessarily consider themselves victims of linguistic imperialism. Danet and Herring (2007:19), for example, reported on studies of Web-based chats in which 'English dominated, generally in a non-conflictual manner. Non-English speakers, being generally bilingual, were willing to switch to English even in set-

tings where the majority of the users were non-English speaking'. In the same volume, a research study carried out by Durham (2007) showed that English was the preferred choice for communication between members of a Pan-Swiss medical student organisation.

The widespread use of English online, however, may not be without consequences, as it strengthens the global position of English, online and offline, 'with the resulting acceleration of the global spread of English' (Danet and Herring, 2007:22). Another possible consequence is suggested by Warschauer *et al.* (2010:491), who indicate that the use of English as the lingua franca of cyberspace could be part of a process of internationalisation of the English language itself, as:

> '...by simultaneously facilitating daily communication in English by hundreds of millions of non-native-speakers around the world, this trend also calls into question who controls English and sets its standards.'

In the same article, Warschauer *et al.* argue that the highly colloquial, informal and hybrid forms of English employed in online interactions also challenge the norms of standard English.

Leaving aside normative considerations, the following cannot be denied: on the one hand, CMC in English might have influenced the linguistic features of CMC in other languages (through the borrowing of computer-related terms, e-grammar conventions, the flouting of orthographic norms, and typographic playfulness, see Herring, 2012); on the other hand, CMC's contexts are increasing and differentiating in terms of languages and socio-cultural configurations (Macfadyen, 2006b) and English does not dominate in every multilingual situation. Interestingly, Danet and Herring (2007:13) observe that 'There is nothing to indicate that the adaptations found [in Swedish] are significantly different [from] online adaptations [in] English or French'. In other words the ways in which human language adapts to the medium appear to be very similar the world over.

The existence of a single online speech community is a controversial issue in CMC research, and current studies more frequently speak of the Web as the virtual environment of many different speech communities, as the various types of CMC are now part of the repertoire of available modes of communication (Murray, 2000:404). Sociolinguistics posits that 'people in regular contact with one another tend to share more linguistic features, and tend to borrow more features of each others' language varieties, even in situations where those varieties are different languages.' (Paolillo, 1999:1) In general, CMC research has shown that participants in online environments bring with them expectations for interaction which were acquired in other online communities or in face-to-face contexts and tend to develop their own rules and routines (Herring, 2010), which are subsequently accepted by newcomers. Another finding of CMC studies concerns self-representation in cyberspace. Macfadyen (2006:472) reports that online identities may be 'multiple, fluid, manipulated and/or may have little to do with the "real selves" of the persons behind them'. CMC brings with it disembodiment and the possibility to create multiple virtual selves, which may have a positive effect on L2 production.

In sum, especially when used for recreational functions, CMC is characterised by informality, non-standard orthography, typographic innovations and adaptation to the technological constraints. In addition, conventions originally noted in English CMC have been found to be used cross-linguistically. The status of CMC as a new variety of English, however, is still an object of debate. The next section deals in greater detail with the distinctive linguistic features of CMC in English, including those found to be also characteristic of CMC produced by non-natives and by speakers of other languages.

### 2.6.2 Linguistic Features of English Text-based CMC

This section is an overview of the linguistic features of English CMC. The choice of the linguistic features described is motivated by their appearing, in various

degrees, in the corpus of CMC collected for the present study. In particular, the following review focusses on findings about grammar, orthographic and typographic conventions, and about its essentially visual nature, epitomised by the extensive use of symbols such as emoticons. The present section includes a brief account of the spoken versus written debate which has animated much of the research on CMC.

Over the past two decades, CMC has emerged as an important new communication modality. Its main features, lower levels of formality and an orientation toward interpersonal interaction (Danet and Herring, 2007), have attracted the interest of multiple disciplines such as cultural studies, intercultural studies, linguistics, sociology, education, learning technologies, and so on (Macfadyen, 2006b). Linguistic studies of CMC have concentrated on a variety of aspects, from discourse-level features, to *e-grammar*, to lexical and typographic innovations.

The term *e-grammar* was initially used to imply the existence of a CMC grammar; however, more recently, scholars have claimed that 'electronic language, as a new and still emergent phenomenon, has not yet had time (nor attained the requisite social status) to become formalized in "rules;" rather, it exhibits patterns that vary according to technological and situational contexts' (Herring, 2011). At the beginning, CMC's distinctive features, like subject pronoun deletion, abbreviation, typos and mixed case, were explained with economy of effort and time constraints. As we saw in Section 2.6.1, however, they originated and spread in the informal, conversational CMC modes, and reflect their essentially recreational function and visual nature.

Further evidence for this interpretation comes from CMC's use of emoticons. Dresner and Herring (2010), for example, argue that emoticons are used as indicators of emotional and non-emotional meanings mapped onto facial expressions, but also of illocutionary force. As for non-standard use of capitalisation and punctuation, researchers cite the influence of SMS language, evident in the sub-

stitution of words or parts of words with numbers or other letters. Once again, the use of these features has multiple functions: saving keystrokes, playfulness, and indicating social or group identity by attempting to imitate the vernacular style (see, for example, the use of *dunno* for *don't know*) (Herring, 2012). The syntax of computer-mediated English has been described as simplified and fragmented. In particular, texts exhibit elision of subject pronouns and articles, and the syntactic function of punctuation marks is mostly neglected.

A criticism that has been levelled at CMC linguistic research is that the data analysed mostly come from case studies and that, to date, large, representative corpora of CMC have not been collected. Beisswenger and Storrer (2008:306), for example, observes that 'up to now, many assumptions about the Internet's impact on language change have been based upon small datasets and a lot of intuition'. On the one hand, CMC is underrepresented in large online corpora such as the BNC or the Bank of English because their creation predates it; on the other hand, as Gong (2005) and Beisswenger and Storrer (2008) pointed out, CMC research has tended to rely on project-related corpora of raw data.

The main problems with creating corpora of CMC reside in text selection and extraction, privacy and copyright issues, as well as in the linguistic and typographic features of the texts themselves, which make them incompatible with automatic analysis and annotation using available morphological analysers or lemmatisation tools. These practical issues were encountered in the electronic analysis of the asynchronous chat corpus used for the present study, and they are discussed in detail in Section 3.3.4.

Another reason for the limited inclusion of CMC in large corpora of English, mentioned by Beisswenger and Storrer (2008:305), is the unclear status of CMC with regard to the spoken–written dichotomy. Collot and Belmore's (1996) analysis of exchanges on electronic bulletin boards led them to consider *Electronic Language* a new variety of English, a hybrid between speech and writing. Similarly, Crystal (2006:5) coined the word *Netspeak* to refer to a new language

variety, noting that 'many of the expectations and practices which we associate with spoken and written language... no longer obtain' (Crystal, 2006:5). In actual fact, over the last decade, English-based research has shown that the various CMC modes tend to display both speech-like and writing-like features, as well as digital ones. The blending of speech and writing is no novelty in language studies. As noted by Gong (2005), in the past, other audiovisual technologies, such as radio and television for example, contributed to the creation of hybrid blends. Whether the distinctive feature of CMC is its speech-like quality is still a matter of debate; some researchers advocate describing CMC genres in their own terms, rather than in terms of how they differ from speech or writing (see Goutsos, 2005). Herring (2010:3), for example, observes that CMC, both in its synchronous and asynchronous modes, exhibits both orality and the interactive and social dimensions associated with face-to-face communication.

Being a global communication medium, CMC has been considered the perfect candidate for language learning, as it provides learners with authentic practice in real-life communicative contexts. Computer Aided Language Learning (CALL) studies have increasingly integrated CMC, and a new branch of language teaching called NBLT (Network-Based Language Teaching) has evolved. NBLT employs computers connected in either local or global networks, allowing one-to-one, one-to-many, and many-to-many communication for pedagogical purposes (Kern *et al.*, 2008). Despite concerns that the type of communication learners practise online might not coincide with that traditionally considered as the object of foreign language learning (Kramsch and Thorne, 2002:83), since the early 2000s teachers and researchers have used CMC in different learning contexts and the following section reports on studies of learner language produced by means of computer.

### 2.6.3 Studies of Learner CMC English

The earliest studies of the effect of CMC on learner production date back to the 1990s, when the preoccupation of researchers was to find a place for CMC in

the language classroom. As a result, its applications and implications for language learners have been extensively researched. According to CALL and NBLT research, the advantages of computer-mediated language production for learners are manifold. First of all, it provides increased opportunities for communicative practice; in addition, potential benefits include: enhanced motivation, greater involvement in language learning, reduction of anxiety, promotion of learner autonomy, and improvement of both receptive and productive language kills. The studies reviewed below are just a sample of many publications that deal with CMC by non-native-speakers for language learning purposes. A considerable number of the early studies were carried out in the United States, and they report on CMC in languages other than English.

In text-based CMC, the lack of the phonetic channel is considered highly beneficial, as pronunciation may be a hindrance to effective communication at the early stages of language learning. Moreover, as typing takes longer than uttering sentences, if they wish, learners can profit from extended processing times and extra time for editing. CMC, in fact, is conceived both phonetically and visually, and so mistakes can potentially be spotted and self-corrected before sending off the text (as noted, *inter alia*, by Dresner, 2005; Kern *et al.*, 2008 and Sauro and Smith, 2010). Sauro and Smith's (2010) study of chats by learners of German as a foreign language, for example, used screen capture software and found that learners monitor their output and self-correct before hitting the return key, and so appear to use the increased time allowed by the chat for production that is more linguistically complex than that produced in face-to-face interaction.

In terms of motivation and involvement, studies indicate that CMC influences attitudes towards using the L2, even though, as Thorne and Black (2007) point out, there is a cultural side to the engagement in CMC. Studies of Telecollaboration, that is, international partnerships that link groups of learners in online discussions, were explored in Kötter (2003), Thorne and Black (2007), and Kern *et al.* (2008). The researchers found cultural attitudes influenced motivation

and involvement. On the one hand, such partnerships multiply communicative opportunities to use the language in meaningful tasks; on the other hand, CMC modes bring with them different expectations and behaviours which are rooted in cultures-of-use, which may result in misunderstandings between the two cultures (as demonstrated in Kramsch and Thorne, 2002 and Thorne and Black, 2007).

Attitudes towards CMC have been explored by sociolinguistic studies in order to find out 'how learners interpret and construct meaning online across culturally situated contexts' (Kern *et al.*, 2008:283). This line of research has taken a great number of directions, from new literacies to studies of roles and identities in institutional contexts, since nowadays CMC pervades work, education and interpersonal communication.

Sauro and Smith's (2010) study exemplifies the kind of interest Second Language Acquisition takes in CMC. SLA scholars have focussed on two lines of research: comparisons between online and face-to-face interaction in terms of negotiation of meaning and noticing, and studies of the effectiveness of CMC in promoting transfer of language proficiency from online environments to real-world ones. As Kern *et al.* (2008) underline, the interest in CMC was fuelled by the ease with which online interactional data can be collected from logs of Internet chats. Both Second Language Acquisition and sociolinguistic-based studies have mainly researched discourse written by post-secondary foreign language learners in both asynchronous and synchronous environments.

Overall, scholars agree that there is a degree of transfer from online environments to real-world ones, as the processes and outcomes of language learning coincide in the two worlds. Abrams (2003), for example, analysed different groups of learners of German in order to test the hypothesis that synchronous CMC has a positive effect on oral performance. In general, SLA views CMC as a form of pre-speech and Abrams found that the increase of production was strictly connected to differing attitudes towards synchronous and asynchronous CMC modes. However, it should be added that, as discussed in the previous section, although

it may be conversational in function and linguistic register, CMC cannot simply be considered speech written down.

While there is an increasing interest in collecting corpora of CMC in EFL contexts, there are very few studies available to date and researchers have mainly been interested in issues of negotiation of meaning and learner language pragmatics. Possible reasons for the lack of corpora of learner CMC could be related to technical or copyright issues, or to the fact that it is still unexplored territory for learner corpus linguistics. One recent study about learner corpora of CMC, Foss (2009), relates the compilation and analysis of JLEBC, a corpus of Japanese Learner English Blogs, which consists of 8,858 blog entries by 654 low/intermediate Japanese learners of English. Learners blogs were compared to native-speaker writing from the BNC in terms of high frequency vocabulary use. By examining the 300 most frequent words in each corpus Foss' study reveals a mix of overuse and underuse of high frequency vocabulary, which is mainly attributed to L1 interference and the limited active vocabulary of the learners.

Another recent learner CMC corpus study is underway at the Mid-Sweden University, where the McCALL, a corpus of Computer-Assisted Language Learning consisting of discussion board messages, is being assembled. The pilot version of the corpus, called Mini-McCALL, was introduced in Deutschmann *et al.* (2009). The McCALL corpus is still being compiled and studied. Ädel (2011) used it to study rapport-building language in online student communication. For Ädel, online student communication is 'a new type of data that we are very likely to see more of in the future, considering the fast speed at which e-learning in general and CALL applications in particular are spreading.' (Ädel, 2011:2945).

Ädel's words pinpoint the reason why CMC was considered appropriate for collecting learner English for the present study. CMC increasingly represents a real-life communication experience for EFL learners, who may use it today for personal communication and in CALL activities, and tomorrow in their future jobs. As Dörnyei and Ushioda (2009:3) point out, in EFL contexts English is

learnt in order to be able to communicate with a non-specific global community of English-language users, often by means of computer-mediated communication.

Even though scholars maintain there is no single online speech community, the cross-cultural and cross-linguistic similarities observed in empirical studies indicate that CMC brings with it some degree of participation in one of the numerous speech communities of the Web. Not being part of a particular speech community, as we have seen in the previous sections, has been repeatedly given as an explanation of the shortcomings of learner English, especially in terms of phraseology. The analysis of learner language produced by means of CMC could provide additional support to these findings.

An additional advantage of CMC for learner data collection is its being a potentially non-threatening environment for L2 production, possibly resulting in spontaneous, highly motivated production. As we have seen in Section 2.6, CMC brings with it a sort of disembodiment and the possibility to create multiple selves, including, possibly, an English-speaking self. Attitudes towards CMC, communicating in English, feeling part of a larger speech community, L2 possible and future selves and real-life experiences with the L2 and its speech communities all play a part in the language produced by the learners investigated for the present study.

Even though studies of learner language produced by means of CMC are accumulating and CMC is increasingly considered a new type of learner data worthy of investigation, to date no published research study has investigated learner CMC in terms of recurrent sequences. However, as shown by the brief overview above, it is a promising research area. Analysing learner recurrent sequences in this medium may provide new insights into learner English at advanced level, complementing and supporting previous findings on learner English. In addition, comparisons with learner English produced through other media, such as those recorded in the ICLE and LINDSEI corpora, may increase our understanding of the ability of advanced learners to adapt their register to the different genres:

writing, speech and CMC, and inform teaching practice and materials.

Chapter Three focusses on the design and building of the Learner Chat Corpus analysed for the present research. After a review of available corpora of learner English in Sections 3.1 and 3.2, Section 3.3 presents a detailed description of the corpus design, comparability and data collection phases. The chapter ends with an overview of the main features of the learner texts included in the corpus and of the learner profiles collected by means of a survey (Sections 3.3.5 and 3.3.6).

# Chapter Three

# Learner Chat Corpus: Design and Building

After the outline of the research aims and methodology in Chapter One and the overview of the relevant literature in Chapter Two, Chapter Three focusses on the design, collection and building of the Learner Chat Corpus (hereafter LCC). This will be followed by an analysis of the data and comparisons with other corpora of learner English in Chapters Four and Five.

In order to contextualise LCC and understand the decisions behind its design, Sections 3.1 and 3.2 review existing learner corpora of both writing and speech collected in various learning contexts throughout the world. In particular, Section 3.1 focusses on the International Corpus of Learner English (ICLE) collected by a research group headed by Sylviane Granger at the University of Louvain, Belgium, because its design was followed closely in the learner English corpus collected for the present work. Section 3.2 presents a detailed review of other learner corpora collected in various countries and institutional contexts.

Section 3.3 is an account of LCC in terms of design and comparability with ICLE, followed by a report on the data collection and corpus building phases, and a detailed description of the features of the corpus and of the learners under observation.

## 3.1 The International Corpus of Learner English (ICLE)

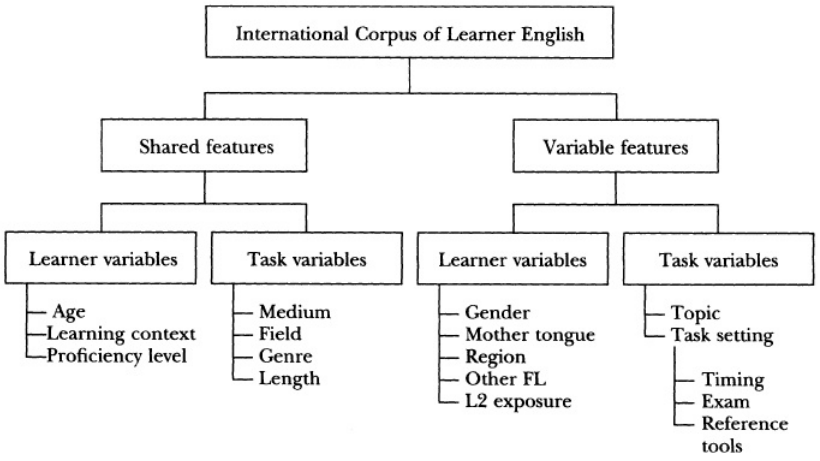
This Section provides a detailed account of the International Corpus of Learner English (ICLE). In addition to being the most comprehensive learner English corpus in terms of language backgrounds, ICLE has been available on CD-Rom since 2002 and has provided empirical data for most of the research on learner English carried out to date. As the Learner Chat Corpus (LCC) used for the present study was designed to be comparable with ICLE, the following paragraphs provide a description of the design criteria and the structure of the corpus.

Sylviane Granger and her team at Louvain started the first major project involving the creation of a corpus of learner English in 1990. According to Granger (1998:6), the advantage of collecting a large corpus of learner English from multiple L1 backgrounds would be that they would gain unprecedented access to the learners' total interlanguage and, therefore, provide generalisable results across L1s. The outcome of their work was a 2.5 million-word corpus of advanced learners' written English from eleven different L1 backgrounds, collected by a team of seventeen universities. ICLE's design criteria are recorded in Granger *et al.* (2002b).

Granger (2004b:125) reported that the main difficulties were, firstly, that learner data was not so readily available for collection and, secondly, that a large number of variables influence L2 production. Learner language is highly heterogeneous in nature; strict control criteria were therefore set, and corpus collectors at the various institutions were learner and task variables. Figure 3.1 below illustrates the design of ICLE and shows shared and variable features both in terms of learners and task.

As can be seen, ICLE learners share the following features: age, learning context and proficiency level; while variable features include gender, mother tongue, region, the knowledge of other foreign languages and L2 exposure. As for tasks,

Figure 3.1: ICLE Learner and Task Features




the texts collected in ICLE share medium, field, genre and length; but they vary in terms of topic and task setting.

The disadvantage of collecting samples from a wide variety of L1 contexts is that learner profiles never match fully and, as a result, shared features show marked national differences. Learners' age is a case in point, learners are between the ages of 21 and 27, but Swedish learners are generally much older than the Bulgarian or Dutch ones. In order to ensure a homogeneous learning context, however, all the students had to be enrolled in a Languages Faculty, with a specialisation in English as a foreign language.

Language proficiency was defined by an external criterion: learners had to be undergraduates in their third or fourth year of university. Granger chose an external criterion because controlling language proficiency across different institutions would not be feasible. First of all, European universities have different requirements in terms of English language proficiency; secondly, there are no universally accepted standards of evaluation, as language proficiency depends on a multiplicity of factors and the Common European Framework of Reference was not yet used extensively.

It is not surprising, therefore, that studies analysing the national subcorpora of ICLE found considerable cross-corpus variability, depending on the proficiency measure analysed. Tono (2003), for example, notes that although the Japanese-

speaking learner group fulfilled the criteria set for ICLE, their proficiency was noticeably inferior to that of their European counterparts. As a matter of fact, as reviewed in 2.5.3, studies that draw their learner data from ICLE tend to focus on one or two subcorpora rather than use the corpus in its entirety. Over the years, however, the importance of ICLE as a representative corpus of Learner English has not diminished: it is still employed as a source of empirical data by a considerable number of research studies all over the world, and its design has been widely imitated.

As for composition, the learner texts included in ICLE are essays produced in academic contexts, mostly argumentative essays (91%), and literature exam papers (with controlled percentages). Text length is between 500 and 1000 words and essays and exam papers cover a range of topics. In terms of task setting, ICLE is a mix of timed exam papers written by learners without the aid of reference tools and untimed papers written with the use of reference tools.

A closer look at the composition of ICLE gives further insights into the learner population included in the corpus. 82% of the texts, for example, were written by female students, as gender distribution in the corpus depends on gender distribution at the faculties of languages in Europe and could be controlled only at the expense of corpus size. Similarly, in terms of L2 exposure, ICLE learners exhibit a large degree of variability, with an average of less than four months spent in an English-speaking country. Knowledge of other foreign languages was a more controllable feature. The faculties of languages in Europe require the study of a second foreign language and learners can generally choose between French, German and Spanish. The majority of the learners included in ICLE studied either French or German.

The 2002 version of ICLE includes texts by European learners from eleven different mother tongue backgrounds, each subcorpus containing over 200,000 words. Learner language samples are complemented with learner profiles, collected by means of a questionnaire and included in the CD-Rom in separate,

searchable files. Information about the learners helps researchers to draw general conclusions about advanced learner English, and also to create ad hoc subcorpora defining learner variables, such as Spanish mother tongue learners who speak some English at home, or learners for whom German is the second language and English is the third language (Granger *et al.*, 2002b). In addition, the data about the learners could be used to analyse sociolinguistic aspects and as a basis for developing teaching materials and tools.

The Italian subcorpus, which will be used for comparison in the present study, is composed of 392 essays, amounting to a total number of 227,085 words. The learner texts were collected at a number of Italian universities including Turin, Bergamo, Rome and Milan (Granger *et al.*, 2002b:34). This is reflected in the fact that the corpus is quite heterogeneous. In particular, 198 essays use an article as a starting point, 61 are literary essays and 133 can be considered argumentative essays. The specificities of the Italian subcorpus are particularly relevant to the present study, as the topics and the article used for reference tend to influence the kind of sequences produced by learners, for example in terms of repetitions of collocations and content-related expressions.

ICLEv2, an expanded version of the 2002 corpus, was published in 2009. It contains 3.7 million words of EFL writing. Learners come from 16 mother tongue backgrounds (with the addition of subcomponents for non-European countries such as China, Japan and Turkey). ICLEv2 differs from the first version also by annotation for word forms, lemmas and part-of-speech. Otherwise, the two versions are highly homogeneous and there were no changes in the corpus collection guidelines for learner and task variables.

## 3.2 Other Corpora of Learner English

English learner corpora collected for research purposes can be divided into two main strands: commercial corpora and corpora developed in academic contexts. The first type is collected by ELT publishers and used to compile dictionaries and

write teaching materials; the second type is collected by academic institutions for SLA or ELT research. Examples of commercial corpora are the Cambridge Learner Corpus, a collection of exam scripts by learners taking Cambridge ESOL exams at all levels and the Longman Learners' Corpus, a 10 million word computerised database made up of language samples submitted by students of English from different nationalities and language levels. The main fields of application of learner corpus data are ELT materials, syllabus design and teaching methodology development. However, as they are not usually available to the research community, they have not yet generated academic research outside of the companies which created them. Additional information about these corpora can be found on the publishers' Websites. [1]

Today, the widespread use of computers in schools and universities and the development of Computer Assisted Language Learning (CALL) makes it possible for institutions to collect their own learner data directly on computer. This type of data is particularly relevant to institutions' stakeholders because it provides an insight into current learner results and may become the basis for changes in the curricula and for improving language teaching locally. The main consequence of the availability of learner data and software is that the number of corpora has increased, even if not all existing collections are publicly available and may not have originated a great number of research studies and publications.

The following sections, 3.2.1 and 3.2.2, give an account of the state-of-the-art of Learner Corpora compilation. Collecting, storing and providing availability to corpora online or as CD-ROMs requires a considerable amount of time and resources and this explains why access to many learner corpora can be short-lived and links to the pages where they are stored may be broken and unrecoverable.

[1] Last accessed on 06/10/2012.

### 3.2.1 Corpora of Written Learner English

A number of learner English corpora developed in academic contexts are described by Pravec (2002) and an extensive review of learner corpora around the world has been compiled by Granger and her team and is available on the Centre for English Corpus Linguistics Website[2]. Overall, the design of the earliest learner English corpora collected in academic contexts in Europe, Asia and the US largely follows that of ICLE. The main differences tend to be in terms of proficiency level, age of learners and text types, with some corpora subdividing the data into different levels and ages.

The first learner English corpora developed in Europe were collected in Hungary, Poland and Sweden and they mostly record exam papers and essays. Janus Pannonius University (JPU), for instance, was collected by Jozsef Horvath from 1992 to 1998 and comprises over 300 essays and research papers by advanced Hungarian learners.[3]

The Polish Learners English Corpus (PLEC) was developed between 1996 and 2005 as part of the PELCRA Reference Corpus of Polish. A joint project by the Department of English Language at Lódz University and the Department of Linguistics and Modern English Language at Lancaster University, it collects written work of Polish learners at different proficiency levels. Learners are teenagers and adults and text types vary from Cambridge exam papers to student essays. The corpus also includes 200,000 words of spoken learner English and for some learners the corpus collectors included more than one text, so that they could be the basis of longitudinal studies[4].

The 1,221,265-word Uppsala Student English Corpus (USE) is a collection of essays from the Department of English of Uppsala University, spanning the years 1999-2001. Created for the study of SLA, it collects 1,489 essays written outside

[2]Last accessed on 06/10/2012.

[3]Three subcorpora of JPU are available online for study and can be consulted via the Search blog field. Last accessed on 12/07/2012.

[4]Further information on PLEC can be found on the PELCRA Website. Last accessed on 12/07/2012

the classroom by 440 Swedish university students of English at three different levels. The average essay length is 820 words and essay topics include culture and literature. Like ICLE, the corpus is accompanied by learner profiles with information concerning the learners' first language, their parents' first language, their grades in English, previous studies and exposure to English. The Uppsala University Website has a section dedicated to the project and the corpus is available for use for research and educational purposes.[5]

As can be seen, ICLE set the foundations of classic learner corpora design criteria and most projects share some, if not all, of its features. This is hardly surprising as Granger and her team are indebted to three major corpus linguistics scholars: Jan Aarts, Sidney Greenbaum and Geoffrey Leech (Granger *et al.*, 2002b). Consequently, ICLE's criteria are also behind the creation of the first learner English corpora outside Europe, both in Asia and in the United States.

In Japan, the first major project was started by Yukio Tono (Tokyo University of Foreign Studies). The Japanese EFL Learner corpus (JEFLL) is a collection of compositions written by over 10,000 Japanese learners of English. JEFLL consists of approximately 700,000 words. In this case, the researcher's interest was longitudinal as well as cross-sectional, and composition subjects range from novice to intermediate levels and learners are junior and senior high school students.[6] Tono also started an international project called the International Corpus of Crosslinguistic Interlanguage (ICCI). ICCI is a collection of younger learners' English essays from different countries (China, Taiwan, Hong Kong, Spain, Austria, Poland, Israel), which amount to about half a million words in total. The essays are comparable in design with the JEFLL Corpus. A similar project is CEEAUS (Corpus of English Essays Written by Asian University Students), collected at Kobe University (Ishikawa, 2008) and freely distributed under Creative Commons (CC) license.

[5]USE can be accessed on the Internet from the Oxford Text Archive. Last accessed on 12/07/2012.

[6]JEFLL is freely available online. Last accessed on 12/07/2012.

In China, several projects were devoted to the collection of learner English texts. One of the first was TeleNex (Tsui, 2003), a corpus of secondary-school writing of various types. The Hong Kong University of Science & Technology corpus of learner English was developed by Milton and comprises texts by Cantonese speaking learners of English. Other Asian projects include the Chinese Learner English Corpus (CLEC), 1 million words of English compositions collected from Chinese learners of English with differing levels of proficiency, covering senior secondary school students, English-major, and non-English-major university students in China, and the Seoul National University Korean-speaking English Learner Corpus (SKELC).

Learner English corpora collected in ESL contexts also share features with ICLE. The Lancaster Corpus of Academic Written English (LANCAWE) and the Montclair Electronic Language Database (MELD), for example, were collected in order to study learner academic English but also to create electronic tools such as grammar checkers and student editing aids. [7]

### 3.2.2 Corpora of Spoken and Computer-mediated Learner English

As this overview reveals, most existing learner English corpora have been built with samples of written learner language, as academic essays and exams are the easiest, if not quickest, type of learner production to collect and digitise. However, researchers in various institutions have started collecting samples of spoken learner English as well, despite the technical challenges of recording, transcribing and analysing speech by non-native-speakers, which tend to make this type of corpus collection decidedly more complex and time-consuming.

One of the earliest corpora collecting spoken learner English is the Louvain

[7] LANCAWE, coordinated by Banerjee, is an on-going project aiming to collect a coherent set of academic writing samples from non-native-speakers of English and it is freely available for use in research and teaching. MELD, collected in the United States by Eileen Fitzpatrick and Milton S. Seegmiller, is a collection of English texts written by university students at an advanced level of proficiency from a variety of native language backgrounds.

International Database of Spoken English Interlanguage (LINDSEI), compiled by Granger and her team at the University of Louvain. The project dates back to 1995, and was the result of the collaborative efforts of Gilquin, DeCock, and Granger. LINDSEI includes over 500 informal interviews with higher intermediate to advanced learners of English from eleven L1 backgrounds. The finalised version of the corpus was made available as a CD-Rom in 2010. The corpus is composed of 1,079,681 words (corresponding to 554 interviews) and the number of words per national subcorpus varies, the French subcorpus, for example, is 143,887 words (with an average of 2,878 words per interview), while the Italian one is only 80,047 words (with an average of 1,601 words per interview. The table in Figure 3.2 shows LINDSEI's variables in terms of interview, learner and interviewer.

Figure 3.2: Learner Variables in LINDSEI

| **LINDSEI Variables** | | |
|---|---|---|
| **Interview** | **Learner** | **Interviewer** |
| Genre | Learning context | Gender |
| Duration | Proficiency | L1 |
| Three tasks | Age | FLs |
| Institution | Gender | Status |
| | L1 | |
| | Country | |
| | Other FLs | |
| | Stay in English-speaking country | |

The genre selected for the spoken English samples was the informal interview; learner production during the tasks was recorded and learners were aware of the recording. Interviews had a set duration of fifteen minutes, which were divided into three different tasks: set topic, free discussion and picture description. The topics for the first task included personal experiences and the discussion of a film or a play. The second task consisted in answering questions about various

subjects, such as life at university, travels abroad and hobbies, and the final task was based on the description of four pictures.

Overall, Granger's team took great care to account for all the variables which may affect learner performance (Granger *et al.*, 2002b). In terms of learning context, learners all studied English in a non-English-speaking country. For most of the participating countries, the interviews were recorded in a single institution, exceptions being the Italian, Japanese and Spanish subcorpora.

Corpus compilers realised that there were some differences in proficiency levels, both across subcorpora and within the same subcorpus. In part, these could be explained by the variable degree of exposure to English from country to country. Additional information regarding proficiency levels was included by submitting random samples to a professional rater who assessed them on the basis of the Common European Framework of Reference. Gilquin *et al.* (2010:10) report that 64% of samples were rated as higher intermediate (corresponding to C1 and C2 in the CEFR), with all the extracts from the Italian subcorpus rated as B2 (and lower). In actual facts, therefore, the whole corpus can be described as higher-intermediate to advanced.

Overall, in terms of age, gender, mother tongue, countries and knowledge of other foreign languages, LINDSEI is very similar to ICLE. The main differences stem from the interactivity of the task, which accounts for the inclusion of interviewer variables such as gender, mother tongue, knowledge of foreign languages and status. LINDSEI adds an interesting dimension to learner corpus research in EFL contexts, as it makes it possible to carry out research across four dimensions: the native/non-native dimension and the speech/writing one, combining them in different ways. An incursion into the French subcorpus by Gilquin *et al.* (2010), for example, shows a significant underuse of delexical *make* in writing and a considerable overuse of the same verb in speech, perhaps due to the overextending of its acceptability caused by the pressure of online language production. LINDSEI is discussed in greater detail in Chapter Five, which carries out a comparison

of the recurrent sequences found in the Learner Chat Corpus collected for the present study and those extracted from the Italian subcomponents of ICLE and LINDSEI.

Other projects have tackled the collection of learner speech, among them the Translanguage English Database (TED) Speech, the ISLE Speech Corpus and COREIL, but they are mainly devoted to the study of phonology and intonation. TED includes audio recordings of a wide variety of non-native English speakers presenting academic papers for approximately 15 minutes each. ISLE includes 17 hours, 54 minutes, and 44 seconds of speech data corresponding to approximately 20 minutes of speech per speaker from 23 German and 23 Italian intermediate learners of English. COREIL (2011) is an electronic oral learner corpus designed to study the acquisition of phrasal phonology and intonation in French and English as a foreign language. The Multimedia Adult English Learner Corpus (MAELC), a database of video of classroom interaction and associated written materials collected since 2001 at Portland State University is particularly interesting for the study of classroom interaction in an ESL context.

Lately, the use of computer-mediated communication by learners of English in academic and non-academic contexts has started attracting the attention of research. The widespread use of computers in schools and universities and the development of Computer Assisted Language Learning (CALL) makes it possible for institutions to collect their own learner data directly on computer. This type of data is particularly relevant to institutions' stakeholders. Collecting samples of learners' performance provides an insight into learner results and may become the basis for changes to the curriculum and for improving language teaching locally.

The availability of data has prompted the creation of a few Learner Corpora of CMC. Examples of this trend are the English Taiwan Learner Corpus (TLC) and the Swedish Mini-MaCALL. The English TLC was created through the writing component of a web-based English learning platform called IWiLL. The corpus, which is over 2 million words, includes comments made by teach-

ers in their everyday process of correcting essays online using the IWiLL essay correction interface. Other learner corpora assembled with samples of CMC are the Japanese Learner English Blog Corpus (JLEBC, 2009), started by a private university in Japan, and the Telecollaborative Learner Corpus of English and German Telekorp (2000-2005), developed at Pennsylvania State University. JLEBC is a written learner corpus composed of lower-intermediate learners' blogs and is over 1.5 million words. Telekorp is a bilingual, longitudinal database comprising computer-mediated NS-NNS interactions between 200 Americans and Germans.

The compilation of the Mini-McCALL is an example of this trend in a European context. Recently presented by Deutschmann *et al.* (2009), it is a 1.3-million-word corpus of computer-mediated communication in the context of online English university courses. It consists of written communication in English between students and between students and teachers connected to online English courses offered by the Department of Humanities at Mid-Sweden University. The CMC texts collected include discussion forum messages, e-mail messages, and documents handed in as assignments. Deutschmann *et al.*'s aim is to compile a 10-million-word corpus of computer-assisted language learning to facilitate research into e-learning. The project is still underway and the compilers point out a multiplicity of possible uses, even though, to date, it has constituted the basis of very few research studies.

The learner corpora of CMC reviewed above are evidence that studies of learner electronic English are gaining ground, both for synchronous and asynchronous CMC, even though, as we saw in Section 2.6.3, most are case studies which rely on smaller sets of data collected in the context of CALL.

The next Section explains the rationale behind the Learner Chat Corpus (LCC) design. LCC was created to be comparable with ICLE in terms of learner and task variables as highlighted in Section 3.3.1. Section 3.3.2 is account of task creation; while Sections 3.3.3 and 3.3.4 are a detailed description of the data collection phases and of the compilation of the corpus. The CMC samples collected and the

profiles of the learners who took part are described in Sections 3.3.5 and 3.3.6.

## 3.3 Learner Chat Corpus Design

Discussing the development of linguistic corpora, Sinclair (2005) warns corpus collectors that 'design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.'[8] Accordingly, the following sections of the present chapter provide the explanation of the design of the Learner Chat Corpus (LCC) and of the design features that make it comparable with ICLE, the account of the complex work of assembling the corpus (3.3.4), and the description of the corpus obtained (3.3.5).

Principles and good practices with regards to corpus design were also outlined in Tono (2003), Wynne (2005), McEnery *et al.* (2006) and McEnery and Hardie (2012). As discussed in Section 1.3, these studies guided the choices that lie behind the corpus collected for the present research. In particular, the design of LCC takes into account the considerations expounded in Tono (2003) and is summarised in Table 3.1.

Initially, the main considerations regarded the kind of language the learners had to submit and the type of task that had to be devised in order to collect the language samples. The starting point was the collection of a type of language that had not been investigated in terms of sequences before and computer-mediated communication was the obvious choice. As we saw in Section 2.6.3, CMC is a communication mode that has increasingly attracted the attention of researchers both in the fields of SLA and ELT. Chiefly, it is regarded as a mode that enhances motivation and reduces anxiety, and, undoubtedly, it represents an unprecedented opportunity for learners to practice the L2 in real communicative exchanges. In addition, it should be mentioned that CMC has the advantage of being easier to collect, since computer-based tasks can be carried out anywhere and they are

[8]Sinclair, J. 2005. "Corpus and Text - Basic Principles" in Wynne (2005). Available online. Last accessed 24/06/2012

directly produced in electronic format.

Among the various CMC modes, chats have been shown to be particularly motivating for language learners. For the present study it was deemed important to ask learners to perform a task they would find motivating and that could be a vehicle for the production of spontaneous communication in the L2. Additional language-related features which were considered in the design phase were informal style and familiar topics, mainly because they would be appropriate to the mode selected. The remaining design features were mainly influenced by comparability with ICLE.

Table 3.1: LCC Design features

| **LCC Design Features** | | |
|---|---|---|
| **Language-related** | **Task-related** | **Learner-related** |
| Mode: computer-mediated | Cross-sectional | Age: early twenties |
| Genre: chat | Elicitation: prepared | Motivation: external |
| Style: informal/spoken | References: not allowed | L1 background: Italian |
| Topics: life, leisure, future plans | Time limitation: free | L2 environment: EFL, undergrad. |
| | | L2 proficiency: institutional status |

Comparability with ICLE is discussed in detail in Section 3.3.1. The Sections that follow (3.3.2, 3.3.3, and 3.3.4), instead, document the data collection phases and the building of the Learner Chat Corpus (LCC).

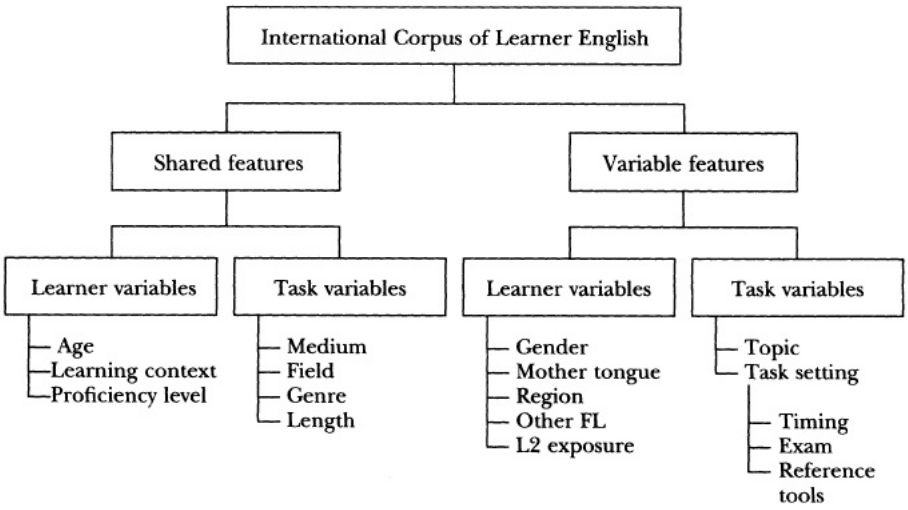
### 3.3.1 Comparability with ICLE

Only a few decades ago, the collection of a large corpus of learner English from a multiplicity of countries was a great endeavour. Even though the ICLE project came at a time when corpus methodology was already established in linguistics, the learner texts contributed were mostly paper-based and had to be converted into digital format. Moreover, controlling all the variables that may affect learner production has an inevitable influence on corpus size. In terms of size, therefore, ICLE is small compared to more recent corpora of learner language. Nowadays, technological advances make it possible to collect and build very large corpora, with much less effort and in shorter periods of time. Larger corpora have multi-

ple advantages: they are more representative of the learner populations and, as highlighted by Sinclair (1991), they enable researchers to generalise their findings. For the present study, therefore, it was deemed important that the size of the corpus should be equal, if not superior, to that of the Italian subcomponent of ICLE (hereafter ICLE_ITA), so that it could be considered as representative of the Italian learner population as ICLE_ITA.

In essence, learner corpus design deals with controlling learner and task variables. Overall, the types of variables that compilers generally seek to control are age, proficiency level and context of learners, and medium, field, genre and length of task (as summarised by Granger in Figure 3.3).

Figure 3.3: ICLE Variables (Granger *et al.*, 2002b)




For the present research study it was deemed essential to collect a learner corpus which would be comparable with ICLE; this meant that learner language had to be collected in a similar learning context and involve a comparable learner population. As a result, learners' age, region and mother tongue correspond, since the task was carried out by students enrolled in the third year of the Faculty of Languages (and therefore in their early twenties) and participants in the chat were undergraduates in a single institution. As is the case with ICLE, learner features which are not shared inter-corpus include gender (even though, statistically, most undergraduates in the Faculty of Languages concerned are female), the other

foreign language studied (the possible L3s are French, German, Spanish, Russian, Arabic and Chinese) and the amount of exposure to English.

As we have already seen in Section 3.1, shared proficiency level is a thorny issue in learner language corpora and some studies argue that it is a fuzzy variable that represents a potential source of error in Learner Corpora. Research is still looking for a reliable method to assign proficiency levels to learner texts. Carlsen (2012), for instance, explored linking learner texts to the Common European Framework of Reference (CEFR). Although it may be a promising path to follow, there are scholars who question the CEFR as the descriptors and their progressions are based on judgments of language teachers and other experts and not based on empirical evidence from L2 learner data (Hulstijn, 2007).

In order to bypass the proficiency issue, for the present research it was decided to define learners' proficiency level by the same external criterion used for ICLE: institutional status (Granger, 2004b:130). Including additional criteria of learner proficiency would have been feasible and was in fact briefly considered. One such criterion, suggested by Tono (2003) in his review of learner corpus design features (see Figure 3.4), is a standard test score. Besides using ad hoc level tests, one could use tests available on the Internet, such as the *Test Your English* online test provided by Cambridge ESOL[9], or could only include learners who hold an internationally recognised qualification. However, the reason why the idea was abandoned is that there is potential variation between learners even in standard tests. Some standard qualification exams, such as FCE or IELTS, for example, tend to assess both receptive and productive skills, but learners' results vary and the productive skills of two learners who hold the same certification may be considerably different. In short, it can be argued that standard tests in themselves do not guarantee uniformity of productive skills any more than English language exams taken at university do. Therefore, and after careful consideration, for the purposes of the present study, standard qualifications were considered as valid an

[9]from the Cambridge ESOL Website. Last accessed on 06/10/2012.

external criterion as any other. Moreover, since the learning context in which the chat corpus was collected was identical to ICLE's, it was decided that it would be more important to adopt the same criterion as ICLE.

In Italy, an increasingly higher number of English language learners take standard examinations for internationally recognised qualifications. At present, however, this is not a requirement for university enrolment. The resulting situation is rather multifaceted, with some learners taking one or more standard examinations during their school years, and other learners having never sat a standard English examination. In terms of internal exams, learners of English in their third year at this institution have passed two English language exams testing the four traditional skills (reading, listening, writing and speaking) and additional skills such as translation into English. As with standard examinations, not all the skills are usually mastered at the same level and productive skills usually lag behind receptive ones.

While the Italian subcomponent of ICLE and LCC share the same features in terms of learners, they differ in terms of task in various ways. Firstly, the chat task was the same for all the learners, while ICLE's essays cover different topics. Secondly, the genre, the electronic chat, was chosen because it would provide insights into how much the medium and genre influence the recurrent sequences produced by the learners. Given this choice, the medium is written, or rather typed, but its linguistic features (discussed in 2.6.2 and 3.3.2) make it a hybrid between speech and writing.

In terms of task timing, assessment and use of reference tools, LCC had no time constraints, and no assessment. The main aim of the corpus data collection was the production of learner language that would be as spontaneous as possible, a language which would not be affected by the psychological pressure typical of speech or exam conditions. Consequently, the length of texts produced during the chats could vary considerably. Participants were asked not to use reference tools, however, since their computers were connected to the Internet there is no

guarantee that they actually did not. Moreover, as the task was not carried out under exam conditions, communication between students cannot be excluded. This is the reason why the use of reference tools could be considered a shared variable. The practicalities of task setting and data collection phases are discussed in more detail in Sections 3.3.2 and 3.3.3.

Table 3.2: Differences between ICLE-IT and LCC

| Variables | | Features | ICLE-IT | LCC |
|---|---|---|---|---|
| Learner | Shared | Age | X | X |
| | | Learning Context | X | X |
| | | Proficiency Level | X | X |
| | | Region | X | X |
| | | Mother Tongue | X | X |
| | Not Shared | Gender | X | X |
| | | Other FL | X | X |
| | | L2 Exposure | X | X |
| Task | Shared | Medium | X | X |
| | | Field | X | X |
| | | Genre | X | X |
| | | Topic | | X |
| | | Length | X | |
| | | Timing | | X |
| | | No assessment | | X |
| | | Reference Tools | | X |
| | Not Shared | Topic | X | |
| | | Timing | X | |
| | | Exams | X | |
| | | Reference Tools | X | |
| | | Length | | X |

During the corpus design phase, it was felt that metadata regarding linguistic background, schooling, L2 exposure and standard tests would complete the picture and, considering the fact that the same information was collected for ICLE, a learner survey was prepared. Learners were asked to complete it after the chat task. The design and results of the survey are presented in detail in Section 3.3.3.

Figure 3.4 shows the learner corpus design considerations expounded by Tono (2003). Although they are not too dissimilar from Granger's (see Figure 3.3), they are particularly interesting at this point because Tono includes internal-affective motivation and attitude. CMC style can be broadly described as personal, infor-

mal and involved (as summarised in Sections 2.6.1 and 2.6.2). The chat mode and the electronic medium are reported to have a strong influence on motivation and attitude (see Thorne and Black, 2007). Motivation and attitude are definitely the most difficult variables to control, however the design of the present study tries to take them into account. The use of chats was deemed interesting in terms of motivation and attitude for at least two reasons: firstly communicating informally with a native-speaker who is their peer was considered to be highly motivating; secondly, young people in their twenties were deemed to be active participants in the global online community. Even though the opposite may be equally true, as personal histories and inclinations vary between learners, the amount of data collected through the chat task is proof that learners perceived it as a motivating task.

Figure 3.4: Learner Corpus Design Considerations Tono (2003)

| Types of feature | | |
|---|---|---|
| language-related | task-related | learner-related |
| mode | data collection | internal-cognitive |
| [written/spoken] | [cross-sectional/longitudinal] | [age/cognitive style] |
| genre | elicitation | internal-affective |
| [letter/diary/fiction/essay] | [spontaneous/prepared] | [motivation/attitude] |
| style | use of references | L1 background |
| [narration/argumentation] | [dictionary/source text] | L2 environment |
| topic | time limitation | [ESL/EFL]/ [level of school] |
| [general/leisure/ etc] | [fixed/free/homework] | L2 proficiency |
| | | [standard test score] |

### 3.3.2 CMC Features and Chat Design

Herring (1996:1) provides the following definition of CMC as 'communication that takes place between human beings via the instrumentality of computers'. CMC has been the object of scholarly research since the 1980s and it has been investigated both in its linguistic and social features. As reported in Sections 2.6 and 2.6.1 most studies focus on text-based CMC, which uses the written word typed on a computer keyboard and visualised on a computer screen.

Traditionally, scholars divided text-based CMC into synchronous and asynchronous modes. The latter were considered closer to the written end of the

written-spoken continuum, and the former were considered more spoken-like (Collot and Belmore, 1996). Asynchronicity was connected to the possibility of planning and revising text, while in synchronous exchanges users, under pressure to type at a conversational pace, were found to produce more non-standard spellings, typos and abbreviations. However, more recent studies (such as Herring, 2010) maintain that the distinction is not so clear-cut, as synchronous and asynchronous modes are found to be overlapping in many cases. In reality, no computer-mediated textual exchange is totally synchronous, since the one-way transmission protocol results in chat exchanges being dependent 'on a rigid succession of messages as they arrive at the computer system' (Yus, 2011:157). A degree of revision is therefore technically possible, and has been shown, even in synchronous CMC. Similarly, asynchronous messages such as emails can be exchanged very rapidly and were frequently found to exhibit the linguistic features of synchronous chats (Herring, 2003 and Gong, 2005).

As discussed in Section 2.6.2, in terms of language, both synchronous and asynchronous CMC have 'topological perceptual characteristics of both spoken and written language' (Dresner, 2005:1) and a number of graphic, orthographic and lexical features that make it stand out as 'a 'hybrid' variety of English' (Herring 1996:28).

While in CMC among native English speakers, synchronous messages are generally shorter, less syntactically complex and lexically less varied than asynchronous ones, due to the temporal constraints of message production and the essentially social nature of the interaction, an interesting reversal takes place in CMC in L2 settings. Sauro and Smith (2010) found that language learners' approach to synchronous chats can be likened to that of native-speakers in asynchronous CMC. Their study of chats by learners of German as a foreign language provides evidence that the textual and technical nature of chats affords longer processing and planning times and, consequently, more complex linguistic output. L2 learners, in fact, have the opportunity to reread and delete portions of

text and can take advantage of longer processing pauses compared to face-to-face conversation. Indeed, the visual nature of CMC, and the fact that words are distanced from the writer and available for inspection and revision, is deemed to have a positive effect on learners' production.

Chat communication is particularly suitable for learner language data collection because of its psychological and linguistic features. The medium has been found to have mainly recreational aims and it is perceived as a psychologically real experience, which decreases inhibition and promotes self-disclosure (Herring 1996), which are particularly relevant variables in learner language production. A text-based chat can be described as a real communicative task that is free of the psychological pressures of both online speech production and exam conditions. Being a communicative task which potentially lowers the learner's affective filter a chat provides the researchers with a perfect means for collecting abundant and spontaneous production, as seen in Section 2.6.3.

Given the fact that in learner CMC features of synchronous and asynchronous communication overlap, for the present study asynchronous CMC was chosen because it is less complex to set up and more controllable than its synchronous equivalent. This type of chat is undoubtedly less realistic and perhaps less challenging than a real chat, but it was considered an effective springboard for learners' output.

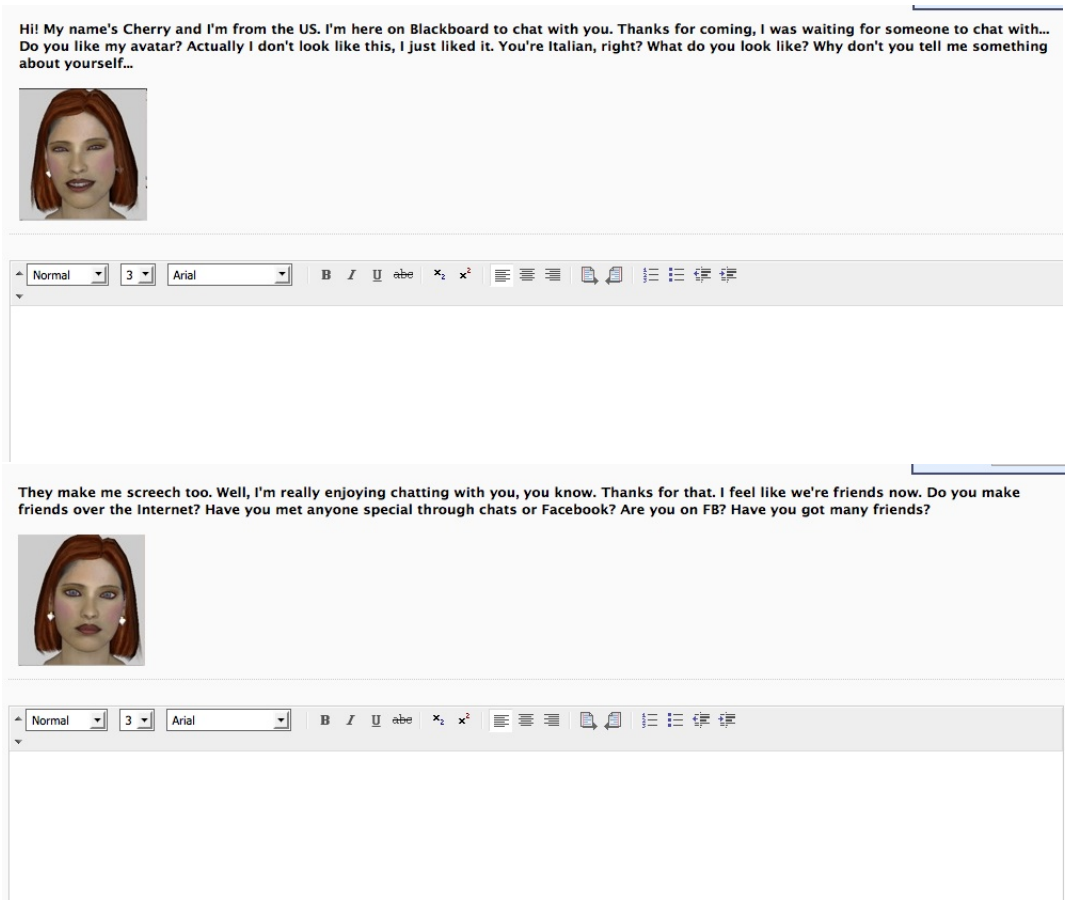
The asynchronous chat created consists of a sequence of stimuli including questions, which resemble conversational turns. The stimuli were assigned to a fictional but potentially real chat-partner, created for the purpose. As most people on chats assume nicks (i.e. aliases) and avatars (i.e. electronic images) in order not to expose their real names and pictures, it was decided to use a fictional character with a nick. The chat text was accompanied by an avatar picture with changing facial expressions (smiling, surprised or puzzled, as the situation required). The stimuli were visualised in succession inside a chat box and learners had another box underneath for their reply, as in a real chat (as shown in Figure

3.5).

After attempting a few characters, complete with a personal and linguistic background, the choice fell on a 22-year-old American female nicknamed Cherry. She was provided with a human-like avatar with red hair and blue eyes. Young Italian students tend to have a fascination with the US more than with any other English-speaking country and it was thought that the learners would relate to the character both in terms of age and interests (travels, nightlife, friendship and FaceBook®). In addition, the character was a native-speaker of English who did not speak any other language so that the learners would be obliged to communicate only in English. The final version of the asynchronous chat consisted in thirteen successive stimuli, written in colloquial English, containing personal questions and showing interest in participants' answers. The stimuli of the chat task are listed in Appendix III.

In terms of software, it was decided to use the university learning platform (BlackBoard®), as it provided a controlled environment which could only be accessed by the chosen participants. The platform records learner data on specific log files and these files could be used to save all the text entered by the learners during the chat. The choice had both advantages and disadvantages. The advantages were that learners knew how to use the platform and they did not need to install new software nor join a new Website; the main disadvantage was that the platform had no asynchronous chat format and therefore another tool had to be adapted to the purpose. This made the visual aspect of the chat screen less realistic than desired. The introduction of the avatar of the chat character did something to attenuate this negative feature, focussing the learners' attention more on the picture and less on the presence of unwanted features. Figure 3.5 shows two screenshots of the chat box including the stimuli, the avatar and the typing space.

Figure 3.5: Asynchronous Chat Boxes




### 3.3.3 Data Collection Phases

Before launching the data collection, there were still a number of practical issues that needed to be addressed. As discussed in 3.3, it was decided to use an external criterion for L2 proficiency level: the target group was restricted to third year students of the Faculty of Languages specialising in English (aged 21 and above).

Data collection was carried out in two successive phases: a pilot collection and the data collection proper. The pilot was launched in 2010. Over 80 learners performed the task and the first data was subsequently extracted and analysed in order to verify the methodology. The results of the pilot were satisfactory both in terms of learner text length, with a total word-count of 56,859 and an average of 646 words per chat, and in terms of learner involvement in the task, as even though participation was voluntary, most learners enrolled in the English linguistics course decided to participate.

The data collection proper took place in 2011, targeting 549 learners enrolled in 4 different English Linguistics third year courses. Students were informed about the task by their professors and two hours in the computer labs were set aside for them for the purpose. The task was voluntary and it was made clear to participants that no assessment was connected to it. The instructions for the task stated that learners should not use any reference materials and that learners should to perform the task in the most natural way. Students who did not do the task in the university computer labs, for lack of time or space, could access it from home and do it from there. Being an untimed task, completion times varied widely. Some learners performed the task in 15 minutes or less, while other learners took 30 to 40 minutes, accessing the task first in the university lab and completing it elsewhere. This variable was not deemed important to control because of unequal typing skills and different degrees of familiarity with the chat medium.

The total number of respondents to the data collection phase was 265. 11 respondents were subsequently eliminated because they did not carry out the chat task, but only accessed it, which included them in the statistics. One respondent was eliminated because she was an exchange student from Australia who did not classify as a learner of English as a foreign language.

### 3.3.4 Corpus Building

The obvious advantage of collecting a corpus by means of a computer is that the data is already in a digital format. However, preparing raw data for analysis by means of a concordancer still involves format conversion. The platform used for the asynchronous chat produced a single output file, a comma separated values file, which included all the student data, the stimuli and the learners' texts (including the 12 failed or invalid attempts). The .csv file was split into 253 separate text files by means of an automatic script that copied and pasted the learners' texts removing everything else.

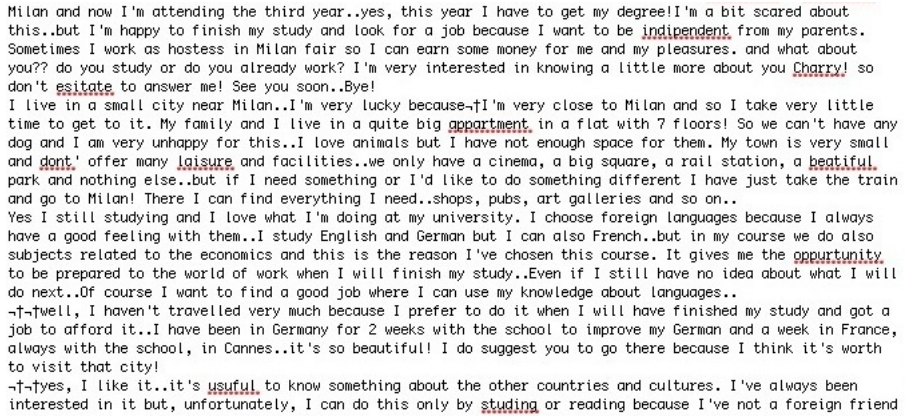
Unfortunately, this process was not as straightforward as one would have wished, as a number of technical problems slowed down the process - from incompatible character sets to formatting tags unwittingly inserted by the students, who used different styles and fonts to embellish or otherwise differentiate their text, highlighting once more the highly visual character of CMC (see Dresner 2005). Learners also made abundant use of other visual CMC features, such as emoticons, acronyms, abbreviations and non standard capitalization, as well as single letter forms (like for example *c u*) borrowed from text-messaging. Some learners employed these features frequently in their chat, thereby showing their familiarity with this medium of communication. Typical chat features appeared also in the creative use of punctuation, which chat users typically employ to imitate prosodic features and indicate pauses, typically to signify that they are thinking.

Another feature of CMC which stood in the way of fast corpus building was typos. Students' typing skills had a strong influence on their production and so did their spelling skills. This is part and parcel of CMC, but since the files had to be analysed by concordancer (Laurence Anthony's AntConc©), amending misspellings was an unavoidable process, as otherwise non standard and wrong spellings of words would influence n-gram results. Misspellings were amended in a copy of the chat files, so that the original text typed by the learner would always be available for future analysis. In order to visually trace each spelling change; the # symbol was introduced before and after each amended spelling.

Figure 3.6 is an example of learner chat viewed in a text editor. As can be seen, there are several misspellings underlined in red. Clearly, some of them are the product of interference from Italian spelling (see *indipendent* and *appartment*), while others look more like classic typos (see *Cherry* vs *Charry*). Most of the times misspellings were quite easy to interpret. One such example is the word *because*, which the learners mistyped in a variety of ways. In other cases, however, the interpretation was doubtful and it was decided to leave the word as it had been

typed by the learner, since a wrong interpretation might influence the content of the learner text.

Figure 3.6: Screenshot of a .txt File from a Text Editor


Milan and now I'm attending the third year..yes, this year I have to get my degree!I'm a bit scared about this..but I'm happy to finish my study and look for a job because I want to be indipendent from my parents. Sometimes I work as hostess in Milan fair so I can earn some money for me and my pleasures. and what about you?? do you study or do you already work? I'm very interested in knowing a little more about you Charry! so don't esitate to answer me! See you soon..Bye!
I live in a small city near Milan..I'm very lucky because¬†I'm very close to Milan and so I take very little time to get to it. My family and I live in a quite big appartment in a flat with 7 floors! So we can't have any dog and I am very unhappy for this..I love animals but I have not enough space for them. My town is very small and dont' offer many laisure and facilities..we only have a cinema, a big square, a rail station, a beatiful park and nothing else..but if I need something or I'd like to do something different I have just take the train and go to Milan! There I can find everything I need..shops, pubs, art galleries and so on..
Yes I still studying and I love what I'm doing at my university. I choose foreign languages because I always have a good feeling with them..I study English and German but I can also French..but in my course we do also subjects related to the economics and this is the reason I've chosen this course. It gives me the oppurtunity to be prepared to the world of work when I will finish my study..Even if I still have no idea about what I will do next..Of course I want to find a good job where I can use my knowledge about languages..
¬†¬†well, I haven't travelled very much because I prefer to do it when I will have finished my study and got a job to afford it..I have been in Germany for 2 weeks with the school to improve my German and a week in France, always with the school, in Cannes..it's so beautiful! I do suggest you to go there because I think it's worth to visit that city!
¬†¬†yes, I like it..it's usuful to know something about the other countries and cultures. I've always been interested in it but, unfortunately, I can do this only by studing or reading because I've not a foreign friend

Foreign words and other non-standard spellings such as mistyped placenames (e.g. *Swizzerland* and *Olland*), words in Italian or languages other than English (e.g. *Fontana di Trevi* and *Sachertorte*), wrong or missing capitalisations (e.g. *george clooneys*) and capitalised words inserted for expressing emphasis (e.g. *you MUST go to Rome*) were left unchanged. The spelling differences between British and American English were ignored (e.g. *favourite* vs *favorite*), as most learners normally do not differentiate between varieties in their writings. Mistakes involving singulars and plurals were also left unchanged, as were compounds written as one word and colloquialisms such as *wanna* and *whazzup*.

Other typographical features had to be changed for technical reasons connected to the use of the concordancing software. These included all Italian diacritics, which were visualised as a tick followed by a cross (i.e. √†) in the concordancer. They had to be replaced with the corresponding non-accented character; for example, *università* was changed to *universita*. Expressive and creative use of punctuation, however, was left unchanged as it can be ignored by AntConc©'s n-gram function. The last issue that required automatic processing of the text files relates to the fact that AntConc© considers words with apostrophes as separate words. In the *Longman Grammar of Spoken and Written English* (1999), Biber *et al.*'s cluster analysis considers contractions as one word, so a script had to be

run which removed all apostrophes.

### 3.3.5 Description of Corpus

The LCC consists of 253 chats by third year undergraduate students enrolled in the Faculty of Language Sciences and Foreign Literatures of Università Cattolica del Sacro Cuore in Milan. It amounts to 205,451 words, after the file cleaning process, and the average number of words per chat is 812. Inter-learner variability is considerably high: the longest chat is 1,947 words, the shortest 80. As discussed in Section 3.3.3, these differences in output can be ascribed to typing skills, and differing attitudes to the task and the chat medium.

The corpus was saved and stored as 253 files in text format (.txt) with no annotation. Part of speech annotation was briefly considered after data collection and some experiments were carried out with two part-of-speech (POS) taggers available online (TreeTagger and CLAWS)[10]. The results of the experiments showed that the learners' errors and features of CMC made POS tagging unreliable. In the end, it was decided to proceed with the analysis keeping the LCC untagged, because of the limited reliability of the tools and because POS tagging was not considered relevant for the aims of the present research.

Apart from typing errors, other types of error annotation were not deemed relevant to the study of recurrent sequences of words for the main reason that error tagging cannot be carried out without a suitable standard for comparison. This is hard in the case of learner corpora, especially when they deal with spoken language or computer-mediated communication. CMC is constantly changing, developing and the result of various types of contaminations (Herring, 1996) and it is impossible to find a standard against which errors could be detected. If we add to that the high incidence of typos, the interactive and online nature of chats

[10]TreeTagger, developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, has been employed to tag texts in different languages (among them German, Italian and French) and is available online. Last accessed on 17/07/2012. CLAWS was developed by UCREL at Lancaster and the latest version is CLAWS4, used to annotate the British National Corpus (BNC). CLAWS4 is available online as a free trial service. Last accessed on 17/07/2012.

and the non-conventional uses of grammar and syntax, it is clear that tagging a corpus of this type would be an extremely time consuming activity, probably destined to failure.

### 3.3.6 Description of Learners

The information on the EFL learners whose samples build LCC was collected in a survey the learners compiled after the chat task. The survey was voluntary, which accounts for the mismatch between the numbers of respondents. The design of the survey follows that devised by Granger for the compilation of ICLE. Questions included regarded learners' mother tongue, the types of school attended and their exposure to English in English-speaking countries. Table 3.7 shows the questions included in the learner survey.

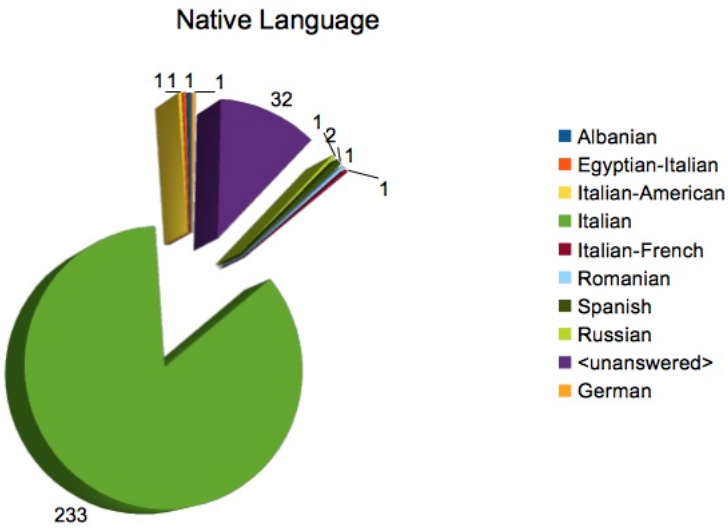
Figure 3.7: Survey Questions

| | |
|---|---|
| 1 | Name of chat completed: |
| 2 | Surname: |
| 3 | First name/s: |
| 4 | Age: |
| 5 | Sex: |
| 6 | Nationality: |
| 7 | Native language: |
| 8 | Father's mother tongue: |
| 9 | Mother's mother tongue: |
| 10 | Language(s) spoken at home (if more than one, please give the average % use of each): |
| 11 | Primary school (name & location): |
| 12 | Secondary school (i.e. scuola media & scuola superiore) (name & location): |
| 13 | Medium of instruction: |
| 14 | Current university studies (curriculum): |
| 15 | Current year of study: |
| 16 | Years of English at school: |
| 17 | Years of English at university: |
| 18 | Stay in an English-speaking country? |
| 19 | Stay in an English-speaking country: WHERE? |
| 20 | Stay in an English-speaking country: WHEN? |
| 21 | Stay in an English-speaking country: HOW LONG? |
| 22 | English language courses abroad? |
| 23 | English language courses abroad? LEVEL? |
| 24 | English Language certifications (i.e. PET, First Certificate,...)? |
| 25 | Other foreign languages in decreasing order of proficiency: |

In terms of language background, the learner survey reveals a very homogeneous learner population. Figure 3.8 shows that out of the 342 students that answered the question about their native language, 233 (96%) said they were Italian speakers. In terms of languages spoken at home, Italian prevails, with

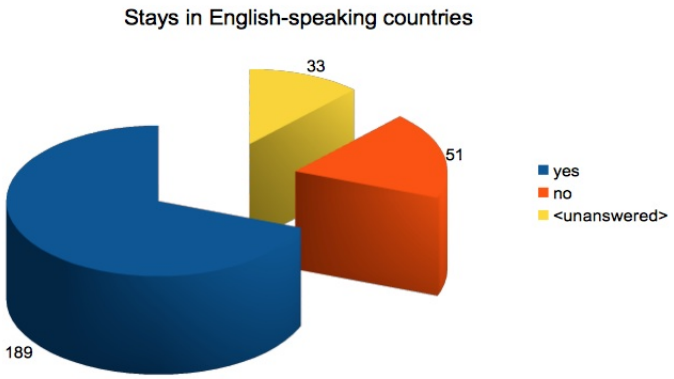
81% of respondents replying that they speak Italian and only 5% declaring they live in bilingual households, with varying percentages of use of the two languages.

Figure 3.8: Learners' Native Languages




As for experience in English-speaking countries, the learners' answers indicate that 69% have travelled at least once to a country where English is spoken, while 18% said they have never had this experience and 12% did not provide an answer. Answers to these questions are visualised in the pie chart in Figure 3.9.

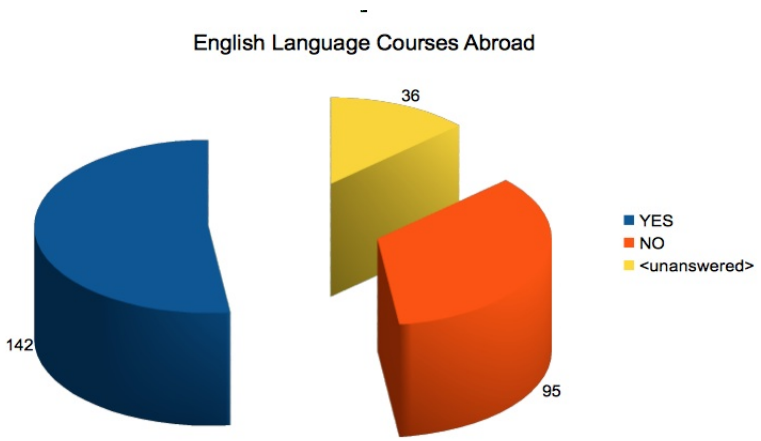
Figure 3.9: Stays in English-speaking countries




Learners' answers to the questions regarding their travel experiences show that, in general, they have travelled widely, with countries visited ranging from Ireland and the UK to the US and Australia (with ten learners including stays in Malta, where English is spoken as an additional rather than a first language), and length of stay between three days and eleven months. In addition to this, 142 learners

said they had taken language courses abroad at various levels and at different stages in their schooling. Ninety-five respondents, however, said that they had never taken English courses abroad. This data is shown in Figure 3.10.

Figure 3.10: Learners' courses abroad




The survey also inquired about English Language Certifications. For this question, learners' answers are too varied to be reported in full. Among the certifications mentioned were Cambridge PETs or FCEs, but 5.7% of participants said they held higher level or ESP certifications such as IELTS, CAE/CPE, TOEFL, BULATS and IPEC. Surprisingly, 33% of respondents said they did not hold any English Language certification.

Chapter Four is devoted to corpus analysis. The data collected in the LCC corpus is analysed in terms of frequencies of recurrent sequences and results are presented and discussed. Chapter Five, instead, provides an analysis and a comparison of the recurrent sequences of words extracted from ICLE_IT, the Italian suborpus of ICLE, and LINDSEI_IT, the Italian component of LINDSEI.

# Chapter Four

# Corpus Analysis

The present chapter deals with the analysis of the specially compiled corpus of learner language LCC. In analysing the corpus, the first step was the automatic extraction of recurring sequences, which was performed by AntConc©. Section 4.1, presents the frequency data regarding the recurrent sequences automatically extracted from the corpus. Data regarding the sequences from 2- to 6-grams are presented in graph format, analysed and discussed. This analysis serves as a preliminary answer to the research question regarding the recurrent word sequences to be found in advanced learner CMC English.

In Section 4.2, the most frequent 3-word sequences from the corpus are classified in terms of structure and function. This part of the analysis brings to light the features of the learners' recurrent sequences and provides an answer to the research questions: what is the structure of the recurrent sequences in LCC, and what are the main functions carried out by these sequences?

The research questions regarding register differences are addressed in Chapter Five, where the data from LCC will be compared to two other corpora of learner English, one written, ICLE-IT, and one spoken, LINDSEI-IT.

## 4.1 Quantitative Analysis

The first type of analysis carried out on the LCC was a quantitative one. As discussed in 1.3, N-grams were extracted automatically by means of AntConc©'s N-Gram function. In the literature, automatic extraction is generally considered the most appropriate way to find recurrent sequences. In her study of learners' preferred sequences, De Cock (2004:228) observes that 'the results yielded by automatic extraction are a useful and powerful starting point as they arguably lead the researcher to take into consideration a series of frequently used clusters he or she may otherwise have overlooked'.

Of the studies using corpus-driven methodology to analyse recurrent sequences reviewed in Chapter Two, reference is made here to Biber *et al.* (1999, 2004), Biber (2006), Biber and Barbieri (2007) and Biber (2009), which analysed native-speaker speech and writing, and to De Cock (2004), Wei (2009) and Ädel and Erman (2012), which made use of automatic extraction of sequences of words to analyse learner corpora. Occasionally, quantitative data is supplemented by qualitative analysis in order to fully investigate the recurrent sequences and provide a detailed account of their features.

Studies of recurrent sequences generally employ a threshold between 20 and 40 occurrences per million word for large corpora (see Biber *et al.*, 1999), and a threshold between 2 and 10 occurrences per 100,000 for small corpora (see De Cock, 1998). In the present study, a cut-off point was set at raw frequencies higher than 20, in other words, only sequences which occurred more than twenty times were considered. There are two main reasons for this choice: firstly, it was deemed important to have a manageable quantity of sequences that could also be analysed from a qualitative point of view. Secondly, higher thresholds ensure n-grams are distributed across texts and are not merely the result of repetition in a limited number of texts. For instance, conversation, as noted by Biber *et al.*, shows a considerable amount of local repetition because expressions from the immediate context tend to be echoed by speakers.

Since one of the aims of the present study is comparison of frequency data with ICLE and LINDSEI, all frequencies extracted from LCC were normalised per 100,000 words. Both the raw and the normalised frequencies of recurrent word combinations automatically extracted from LCC are presented in table format in Appendix I.

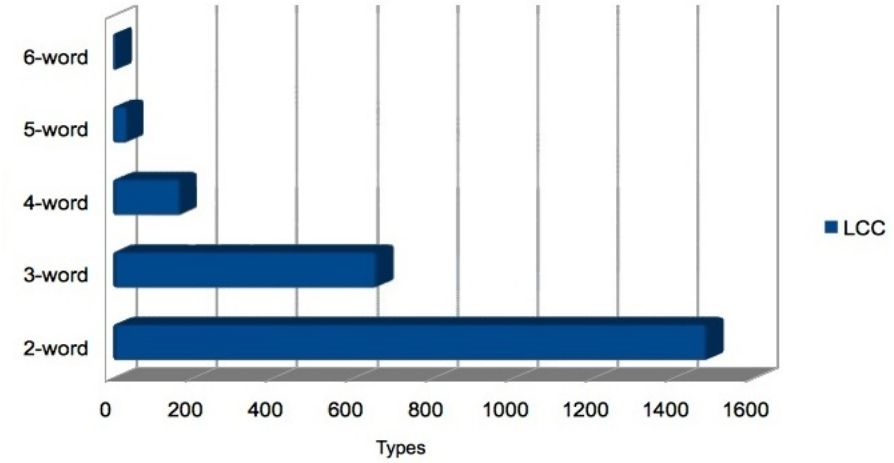
### 4.1.1 Recurrent Sequences

Table 4.1 shows the overall distribution of word sequences in LCC in terms of types and tokens, as extracted by AntConc©'s N-Gram function.[1] Figures 4.1 and 4.2 are graphic representations of the same frequencies.

Table 4.1: Overall frequencies of 2- to 6-word sequences in LCC

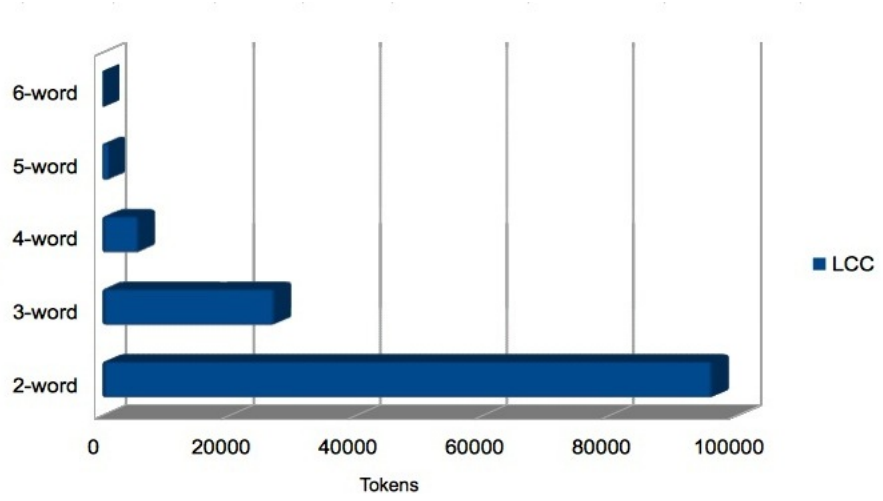
| LCC (Learner Chats Corpus) | | |
|---|---|---|
| Sequence | Types | Tokens |
| 2-words | 1,486 | 96,311 |
| 3-words | 661 | 27,280 |
| 4-words | 170 | 5,799 |
| 5-words | 36 | 1,089 |
| 6-words | 7 | 187 |

Figure 4.1: Frequencies of Sequences of Different Lengths in LCC: Types




[1]For the purpose of the automatic extraction of sequences, all the text in LCC was treated as lower case. As discussed in Section 2.6.1, capitalisation in CMC does not follow the conventions used in written English. In addition, following Biber *et al.* (1999), two-word contracted combinations are counted as one word rather than two. Consequently, contractions in N-grams do not have apostrophes.

Figure 4.2: Frequencies of Sequences of Different Lengths in LCC: Tokens



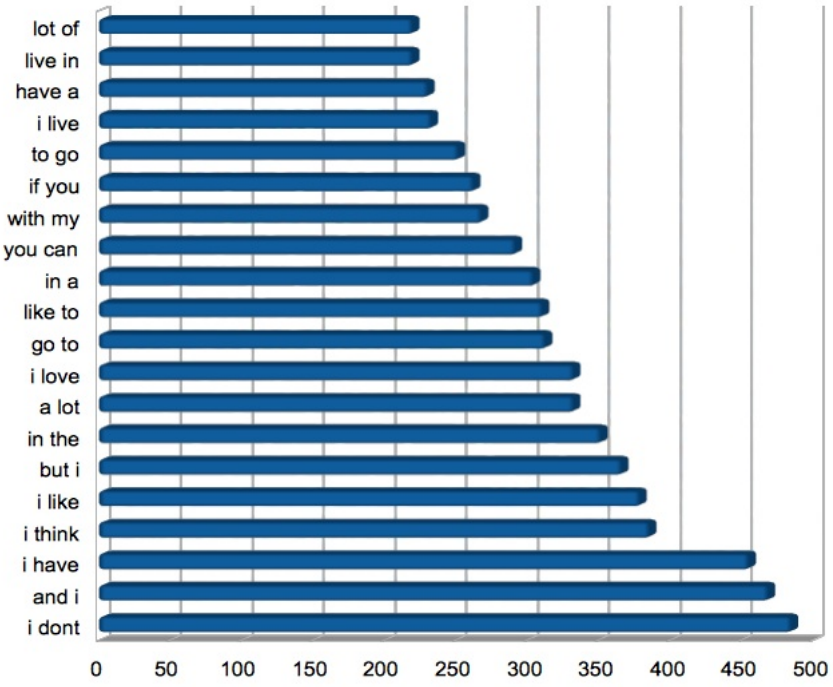

As can be seen, the highest number of types (1,486) corresponds to 2-word combinations. Types decrease considerably for 3-word (661) and 4-word sequences (170); and there are very few 5- and 6- word sequences repeated more than 20 times in the corpus (36 and 7 respectively). In terms of tokens, the overall number of occurrences of all the types, 2-word sequences are over three times as many as 3-word ones, while 3-word sequences are five times as many as 4-word ones. The ratio is similar when considering 4- and 5-word sequences.

In their analysis of lexical bundles in native-speaker speech and writing, Biber *et al.* (1999) found that 3-word lexical bundles are almost ten times as many as 4-word lexical bundles in both conversation and academic prose. A similar ratio was found when comparing frequencies of 4- and 5-word lexical bundles. In the case of the LCC, the quantitative differences are less dramatic. To give an example, the most frequent 3-word sequence, *a lot of*, does not appear even twice as many times as the most frequent 4-word one, *i would like to*. However, it should be borne in mind that corpus size plays a significant role in quantitative findings.

The next sections present the most frequently repeated 2- to 6-word sequences in LCC. The frequency findings can be found in table format in Appendix II.

**2-word sequences**

Figure 4.3: The 20 Most Frequent 2-word Sequences in LCC




The graph in Figure 4.3 shows that the most frequent 2-word sequence includes the subject pronoun *I* and the contracted negative auxiliary *don't*. While no other contracted or non-contracted negative form appears in the top twenty 2-grams, the subject pronoun *I* features in eight different sequences, and the only other subject pronoun is *you*. The most frequent verbs in the learners' sequences are *have*, *think*, *like*, *love*, *go*, *can* and *live* and the only modal auxiliary in the list is *can*. The verbs in learners' most repeated 3-word sequences are either in the base form or in the present tense form. Three sequences include a preposition (*in the, in a*, and *with my*) and three sequences include the conjunctions *and*, *but*, and *if*.

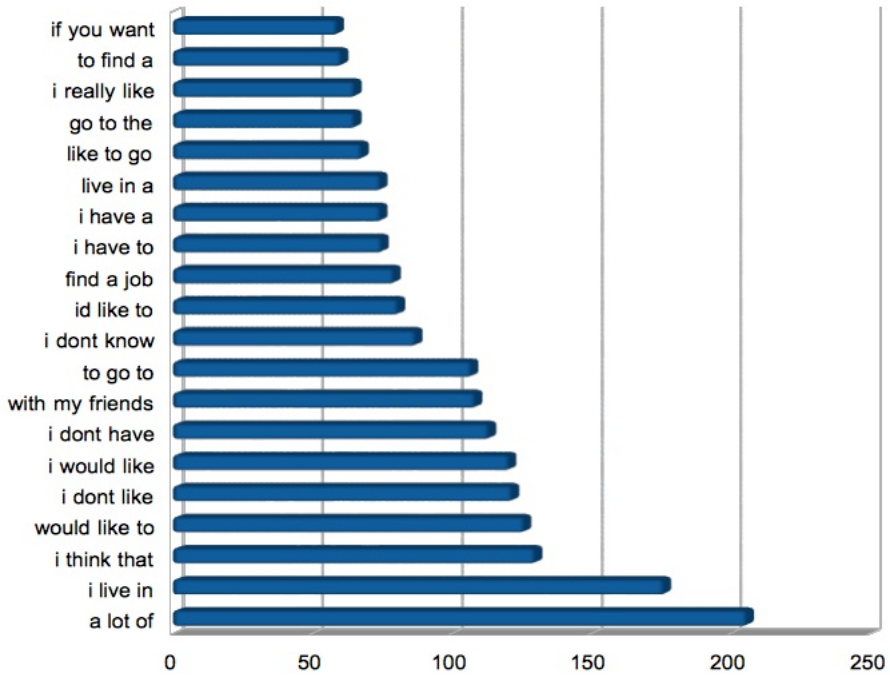
According to Biber *et al.* (1999:994-5), in native-speaker conversation the most common lexical bundles include both *I* and *don't*. In terms of verbs, the most frequently found after *don't* are *know*, *think*, and *want*. Other verbs in high frequency lexical bundles are *say*, *have* and *mean*.

Learner sequences from LCC exhibit a similar prominence of the pronoun *I*. The negative form of the auxiliary *do* appears in the most frequent LCC sequence

and verbs such as *have* and *think* also appear quite high on the frequency list. LCC's highest frequencies, however, also include verbs such as *like*, *love*, *go* and *live*, which do not appear in Biber's most frequent lexical bundles. This difference might be explained by local repetition. Given the size of LCC and the fact that the chats share the same topics, the higher incidence of local repetition may explain the frequencies of these verbs. As the corpora compared in the present study are all small corpora, it is expected that the quantitative analysis of ICLE_IT and LINDSEI_IT (see section 5) will bring to light the same phenomenon.

### 3-word sequences

Figure 4.4: The 20 Most Frequent 3-word Sequences in LCC




As can be seen in Figure 4.4, the 3-word sequence *a lot of* is repeated the most by learners, with nine learners repeating it six or more times in the same chat. In English, the determiner *a lot of* is used to indicate large quantities and it combines with both uncountable and plural countable nouns. According to Biber *et al.*'s (1999) corpus findings, in native-speaker English *a lot of* and *lots of* are characteristic of casual speech. Therefore, while learners may be repeating

this sequence because it eases language processing, given that it can be employed before all types of nouns, the effect of the constant repetition is that it lends the learner language a more spoken, casual quality. As the learners' use of the sequence *a lot of* is particularly interesting from the point of view of language processing and register differences, a closer look at this sequence across the three corpora will be the subject of section 5.3.2.

The 3-word sequences that follow, namely *i live in, i think that*, *would like to*, *i don't like*, *i would like*, *i don't have*, *with my friends*, *to go to* have normalised frequencies higher than 100 per 100,000 words. This means that the repetitiveness of these sequences in the corpus is considerably high, which is quite unusual for context-bound sequences like *with my friends* and *i live in*. Again, this might be accounted by the fact that LCC samples share the same topics.

In terms of composition, the most frequent 3-word sequences are dominated by subject pronouns and verb phrases; in fact, only two sequences (*a lot of* and *with my friends*) include neither a pronoun, nor a verb phrase. Moreover, only three verb phrases have a negative form, which indicates that the affirmative dominates the learners' most frequent verb phrases. In terms of modal auxiliaries, the only one repeatedly used by learners in 3-word sequences is *would*, employed in the expression *would like*.

Only four 3-word sequences have contracted forms, three of which include the negative auxiliary *don't*. Interestingly, the non-contracted form *I would like* is more frequent than the corresponding contracted form *I'd like*. This phenomenon has also been observed in Indian English (ICE-India) in a study by Götz and Schilk (2011:92). They explain the preference with the 'more formal nature of ESL-varieties', which seem to be characterised by a writing-oriented overall flavour.

Contracted forms are considered typical of the spoken register; corpus data from Biber *et al.* (1999), for example, show that they are prevalent in conversation. As LCC is a corpus of chats, contracted forms would be expected to

be more frequent than non contracted ones. As a matter of fact, writing non-contracted forms requires more keystrokes, therefore it is rather surprising to find more non-contracted forms than contracted ones. In order to delve into this apparent contradiction, a qualitative analysis of *would like* sequences was carried out. An analysis of the concordance plot, that is, the distribution in the texts of *i would like* and *i'd like to,* shows that the learners tend to employ one or other of the two sequences. Seventy-eight learners repeat one of the two sequences more than once in their chat (sometimes up to five or six times), and only 15% of them mix *i would like* and *i'd like to* in the same chat. A further concordance search reveals that in only twenty-six instances the learners interrupt the sequence with a pre-modifying adverb (namely *also*, *simply* and *really*). Twenty-six interruptions are very few, given that the uninterrupted non-contracted and contracted sequences are repeated 248 and 174 times, respectively, in the whole corpus.

The learners' repeated use of the same sequence, and their reluctance to insert adverbs between *'d/would* and *like* may be interpreted as further evidence that learners cling on to specific sequences and use them as preassembled chunks of language. As was discussed in the literature review, chunking is widely considered a feature of spontaneous conversation which reduces processing effort. This interpretation is consistent with previous findings on learner English and with findings on other learner sequences from the present study related in Sections 5.2 and 5.3.2.

A noticeable feature of the 3-word sequences produced by learners in the chat task is that they tend to be complete, rather than incomplete, structural units. In general, most sequences contain both the subject and the full verb phrase. Some sequences also include a determiner or a preposition, which provide the beginning of the subsequent noun or prepositional phrase, as in *i live in* and *i have a.*

Evidence from Biber *et al.* (1999) shows that the most frequent 3-word lexical bundles in native-speaker English include personal pronouns and extended verb phrases, but they tend not to be complete structural units. With respect to

personal pronouns and verb phrases, LCC's 3-word sequences are not too dissimilar from native-speaker sequences. However, the structural features of learner sequences seem to differ considerably from native-speaker ones. In order to account for this peculiar feature of learner English, an in-depth analysis of the structure of 3-word sequences from LCC is carried out in Section 4.2.

Sequences expressing negative personal states and personal stance, such as *i don't like* and *i don't know* and sequences including the verb *think*, are very frequent in native-speaker conversation (Biber *et al.*, 1999:1003-4). These sequences appear among the most frequent in LCC, which might indicate that the language used in the chats by the learners is more influenced by speech than by writing. Previous findings on spoken learner English, such as De Cock (2004), suggested that learners mix spoken features with more formal ones, borrowed from writing. In LCC, there are no sequences from writing in the top twenty 3-grams and learner language seems to be appropriate to the mode of communication. However, quantitative findings need to be supported by qualitative analysis before any conclusions can be safely drawn.

**4-word sequences**

Figure 4.5: The 20 Most Frequent 4-word Sequences in LCC

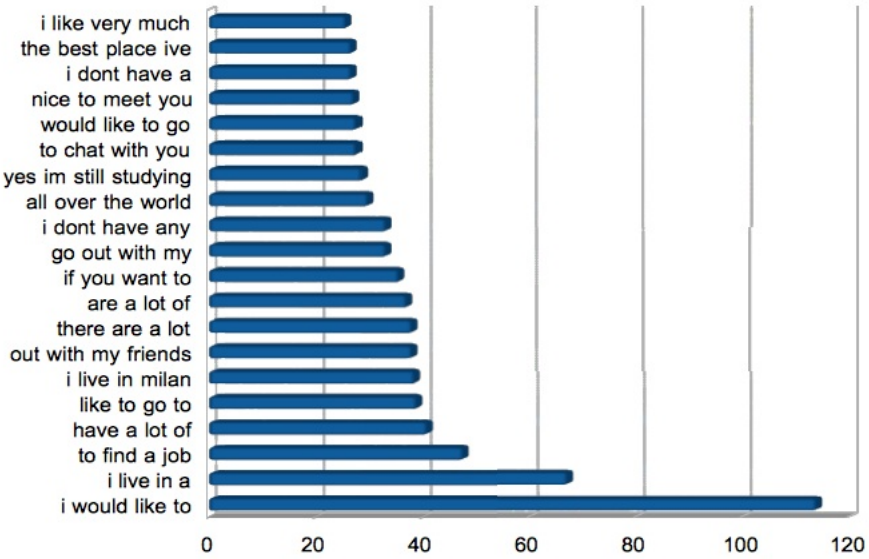


Figure 4.5 shows that the most repeated 4-word sequence, *i would like to*, occurs in the corpus almost twice as many times as the second most frequent sequence. Overall, 4-word sequences are far less frequent than 3-word ones. Moreover, as can be seen in the graph, the most common 4-word sequences include 3-word ones. Examples are: *i would like to*, *i live in a*, *to find a job*, *have a lot of*, *out with my friends*, and so on.

In terms of subject pronouns, *I* is still dominant, it is present in seven out of twenty sequences, while *you* appears only once in subject position (*if you want to*) and in a prepositional phrase (*to chat with you*). Other elements used as subjects are *there* and the noun phrase *the best place*.

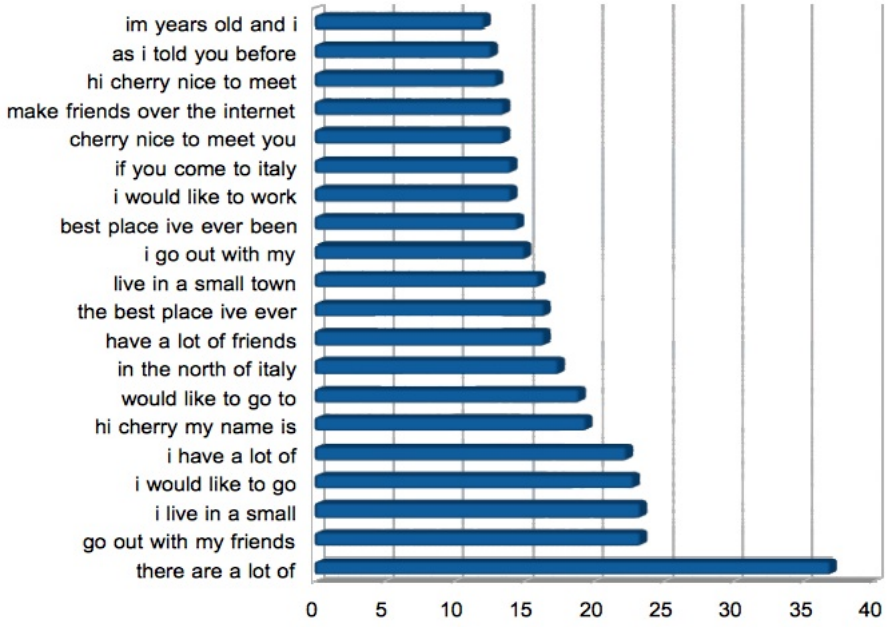
Verb phrases constitute the core element of sixteen 4-word sequences out of twenty. Some sequences include verbs that were not present in the 2- and 3-word sequences, such as *find*, *be*, *study*, *chat*, and *meet*. In terms of tenses, out of a total of eighteen sequences including verbs, five have a non-finite verb phrase and thirteen a finite verb phrase in the present tense. One verb phrase includes the auxiliary of a present perfect tense, *the best place i've*. Moreover, only two 4-word sequences include a negative form, *i don't have any* and *i don't have a*.

Among 4-word sequences are the formulaic expression *nice to meet you* and the quasi-formulaic sequence *all over the world*. Both are context-related expressions, but perhaps the most interesting one is *all over the world*. A concordance search of *world* shows that other learner recurrent sequences including this noun are used even more frequently, namely *in the world* (114 occurrences) and *around the world* (99 occurrences). Similarly to what happens with *would like* sequences, *world* combinations seem to be used by learners as a quasi-formulaic, preassembled expression, and this might be an explanation for their frequency.

In general, 4-word sequences show the same features as 3-word ones. They are mostly made up of full structural units, and they are characterised by verb phrases in the affirmative form, local repetition, and formulaic status.

**5- and 6-word sequences**

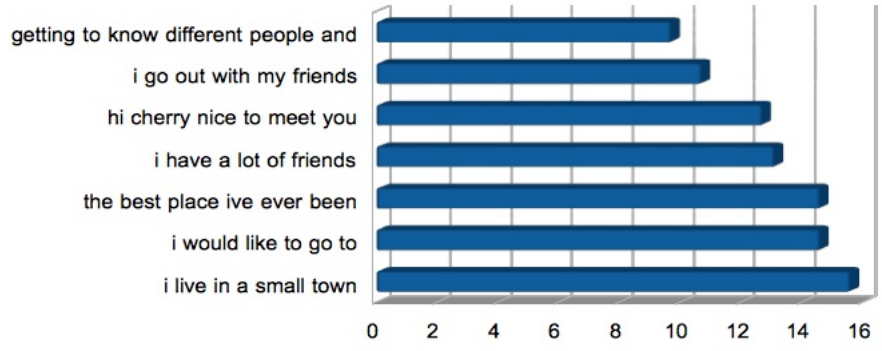
Figure 4.6: The 20 Most Frequent 5-word Sequences in LCC




As reported by most studies of word-sequences, the length and frequency of sequences of words are inversely related: the longer the sequence, the fewer types and tokens can be found in a corpus. For this reason, most studies focus on shorter sequences, especially when analysing fairly small corpora. In line with this, as can be seen in Figure 4.6, 5-word sequences in LCC are quite infrequent, and even the top one, *there are a lot of*, has a normalised frequency of 37. Among the most frequently repeated 5-word sequences is a salutation (*hi cherry*) combined with formulae (*nice to meet you*) and the number of content words, namely nouns, lexical verbs and adjectives, is higher in these sequences than in the 3- and 4-word sequences. In general, 5-word sequences are formulaic, context-dependent and structurally complete.

The 6-word sequences extracted from LCC shown in Figure 4.7 have normalised frequencies between 15.6 and 9.7. In general 6-word and 5-word sequences share similar features. Three of the 6-word sequences are complete clauses, *i live in a small town*, *i have a lot of friends*, and *i go out with my friends*; one is composed of a salutation followed by a formula, *hi cherry nice to meet you*; the remaining

Figure 4.7: The Most Frequent 6-word Sequences in LCC




ones are an extended verb phrase, *i would like to go to*, a dependent clause, *the best place i've ever been* and a non-finite verb phrase followed by a noun phrase, *getting to know different people and*.

According to Biber *et al.* (1999), longer sequences tend to be more context-dependent and include more content words. Data from the *LGSWE* show that 'longer lexical bundles are usually formed through an extension or combination of one or more shorter bundles' (1999:993). This phenomenon can be clearly seen in the LCC data. A comparison across all the sequence lengths shows that, for example, *i have* (935 tokens), the third most frequent 2-word sequence, can be extended to form the 3-word sequence *i have a* (153 tokens), which in turn forms part of 4-word sequence *have a lot of* (84 tokens), 5-word sequence *i have a lot of* (46 tokens) and, finally, 6-word sequence *i have a lot of friends* (27 tokens). A similar extension can be carried out with other 2-word sequences, such as *with my*, and *i live*. In general, the findings regarding the longer sequences indicate that recurrent multi-word expressions in learners' asynchronous chats have a prominently informative focus; they are therefore less interesting in terms of structures and functions and will not be considered in the quantitative analysis.

Overall, the quantitative findings have provided a provisional answer to the first research question of the present study. They indicate that learner English is characterised by the repetitive use of a restricted number of sequences, some of which acquire a quasi-formulaic status. In terms of register, findings show that learner CMC English shares some features of native-speaker conversation.

Since CMC is generally considered closer to speech than writing, this might be considered evidence that learners produce sequences that are consistent with the mode of communication they are employing. As this is one of the claims regarding learner English the present research study set out to investigate, it will be taken up and discussed in Chapter Five in which the sequences from LCC are compared to the most frequent sequences extracted from a corpus of learner writing and a corpus of learner speech.

## 4.2 Patterns and Functions: a Qualitative Analysis

The quantitative analysis carried out by means of a concordancer, related in Section 4.1, revealed that learner English sequences in LCC are limited in number, repetitive, formulaic and context-dependent. These sequences are now analysed in two phases: the first is aimed at determining which patterns are more frequent in learner recurrent sequences while the second classifies learners' recurrent sequences in terms of function. These classifications will provide a fuller picture of learners' use of English in terms of sequences and further insights into their ability to adapt to the mode of communication.

The following analysis focusses exclusively on 3-word sequences. The reason behind this choice is that the 2-word combinations in LCC are too short for classification, while, as discussed in Section 4.1.1, LCC's 4-, 5- and 6-word sequences tend to be structurally complete - some of them being full clauses - and they are closely connected to the chat content. While with very large corpora, such as those used by Biber *et al.* (1999), the automatic retrieval of context-dependent lexical bundles produces a relatively small amount, corpora such as LCC, where the topics are shared by speakers, tend to produce larger amounts of data.

Some corpus-based studies of learner English, such as Baker and Chen (2010) and Ädel and Erman (2012) for example, remove content-dependent sequences

from their frequency counts and ensuing analyses. However, context-dependent bundles also appear in native-speaker English (Biber *et al.* (1999), for example, cite among the most recurrent bundles *put the kettle on* and *have a cup of tea*). These sequences, in fact, are not devoid of interest, as the researcher might gain further insights into the language and the cultural contexts in which it is used.

As the present study aims to uncover the general features of recurrent sequences in learner CMC English, it was decided not to remove context-dependent sequences. On the one hand, the decision is justified by the size of the corpus and by the information exchange focus of the asynchronous chat task. On the other hand, context-dependent sequences are interesting for their own sake, as they provide additional material for cross-corpus comparison.

### 4.2.1 Description of Patterns

The following classification of recurrent sequences by patterns follows that initially developed by Biber *et al.* (1999) for the *LGSWE* and subsequently extended in Biber *et al.* (2004) and Biber (2006). These studies divide lexical bundles into three broad structural types:

- Type 1 lexical bundles include verb phrase fragments and a first, second or third person singular pronoun. Verb phrase fragments can be preceded by connectors and discourse markers. This type also includes yes-no and wh-question fragments.

- Type 2 lexical bundles incorporate dependent clause fragments. These can be to-clauses, wh-clauses, if-clauses, and that-clauses. They are often preceded by a first or second person pronoun.

- Type 3 lexical bundles include noun and prepositional phrase fragments. This type features comparative expressions and noun phrases followed by of-phrase fragments, or other post-modifier fragments.

Table 4.2 shows LCC's top twenty 3-word sequences in order of frequency classified into structural types, broadly following Biber *et al.*'s (2004:381) types. As is customary for this type of analysis, the cut-off point is arbitrary and it was chosen because it is sufficiently high to ensure that the units retrieved have prominence in the corpus. Normalised frequencies (per 100,000 words) are provided for all the classified sequences.

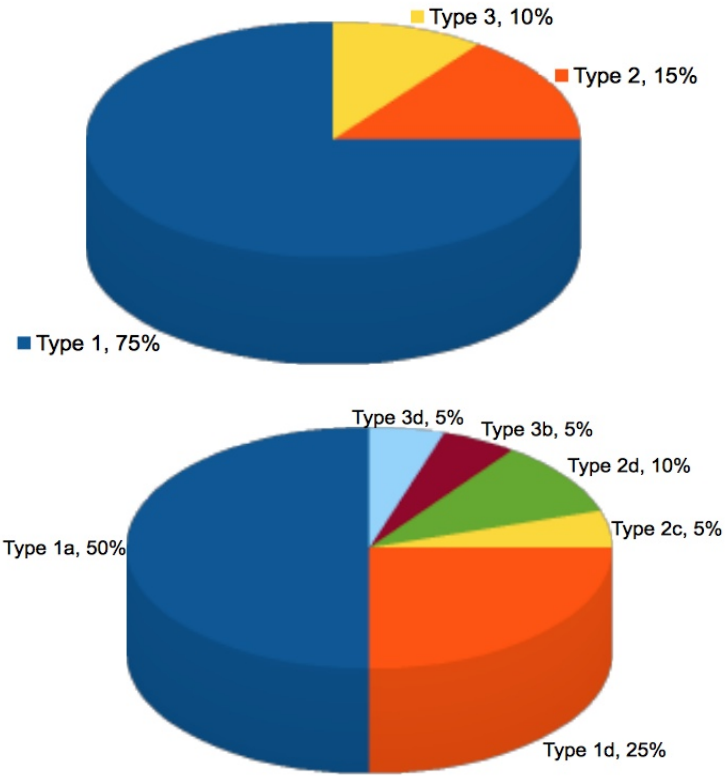
Table 4.2: 3-word Sequences: Structural Types

| Rank | Norm. Freq. | Sequence | Structural Type | |
|---|---|---|---|---|
| 1 | 205.9 | a lot of | 3b | noun phrase with post-mod. |
| 2 | 176.2 | i live in | 1a | 1st person pronoun + VP |
| 3 | 130 | i think that | 1a | 1st person pronoun + VP |
| 4 | 126.1 | would like to | 1d | VP fragment |
| 5 | 121.7 | i dont like | 1a | 1st person pronoun + VP |
| 6 | 120.7 | i would like | 1a | 1st person pronoun + VP |
| 7 | 113.4 | i dont have | 1a | 1st person pronoun + VP |
| 8 | 108.5 | with my friends | 3d | Prepositional phrase exp. |
| 9 | 107.1 | to go to | 2d | to-clause fragment |
| 10 | 87.1 | i dont know | 1a | 1st person pronoun + VP |
| 11 | 80.8 | id like to | 1a | 1st person pronoun + VP |
| 12 | 79.3 | find a job | 1d | VP + |
| 13 | 75 | i have to | 1a | 1st person pronoun + VP |
| 14 | 74.5 | i have a | 1a | 1st person pronoun + VP |
| 15 | 74.5 | live in a | 1d | VP |
| 16 | 67.7 | like to go | 1d | VP |
| 17 | 65.2 | go to the | 1d | VP |
| 18 | 65.2 | i really like | 1a | 1st person pronoun + VP |
| 19 | 60.4 | to find a | 2d | to-clause fragment |
| 20 | 58.9 | if you want | 2c | if-clause fragment |

Figure 4.8 is a visual representation of the same data. As can be seen, Type 1 sequences, first, second or third person singular pronoun followed by a verb phrase fragment, are by far the most common in LCC and they constitute 75% of all the recurrent sequences classified in terms of structures. Type 2, dependent clause fragments, and Type 3, noun and prepositional phrase fragments, only constitute 15% and 10% respectively of the twenty most frequent 3-word sequences. LCC 3-word sequences show a marked prevalence for Type 1a, first person pronoun followed by verb phrase (50%), type 1d, verb phrases (25%), and Type 2d, to-

clause fragments (10%). All the other types are used much less.

Figure 4.8: LCC Sequences: Structures by type




In terms of structural patterns, learner sequences most frequently include personal pronouns and active verb phrases. As was noted before, a feature of the learner sequences in LCC is that they tend to be structurally complete, rather than incomplete. Furthermore, no passive verb phrases or discourse markers appear among the learners' most frequently repeated sequences. These two features are considered typical of the written and the spoken registers respectively. The absence of discourse markers indicates that learners do not use standard multiword expressions to organise their discourse as often as native-speakers do. This indicates that learners organise their thoughts using different linguistic means. This issue is discussed further in Section 4.2.2, which analyses the functions of learners' recurrent sequences.

Question fragments are also noticeably absent and this is rather surprising, given the fact that the chat was an interactive, albeit asynchronous, task. In order to verify the actual presence of questions in the chats, a concordance search

was carried out, using the question mark as a search symbol. The concordancer retrieved 383 questions, a sample of which is shown in Figure 4.9. As can be seen, learners who took part in the chat task did ask questions, but used a wide range of structures to do so. Indeed, many of the questions in LCC present non-standard structures, which are typical of speech. As discussed in Section 2.6, CMC has been described as a mode of communication that closely resembles 'speech written down', which is reflected in LCC in the questions *what else?*, *the best place?*, *mountains or seaside?*, *really?* These questions would seem to indicate that the task was very interactive and that the language produced was highly influenced by features of speech.

Figure 4.9: Sample of Questions from LCC

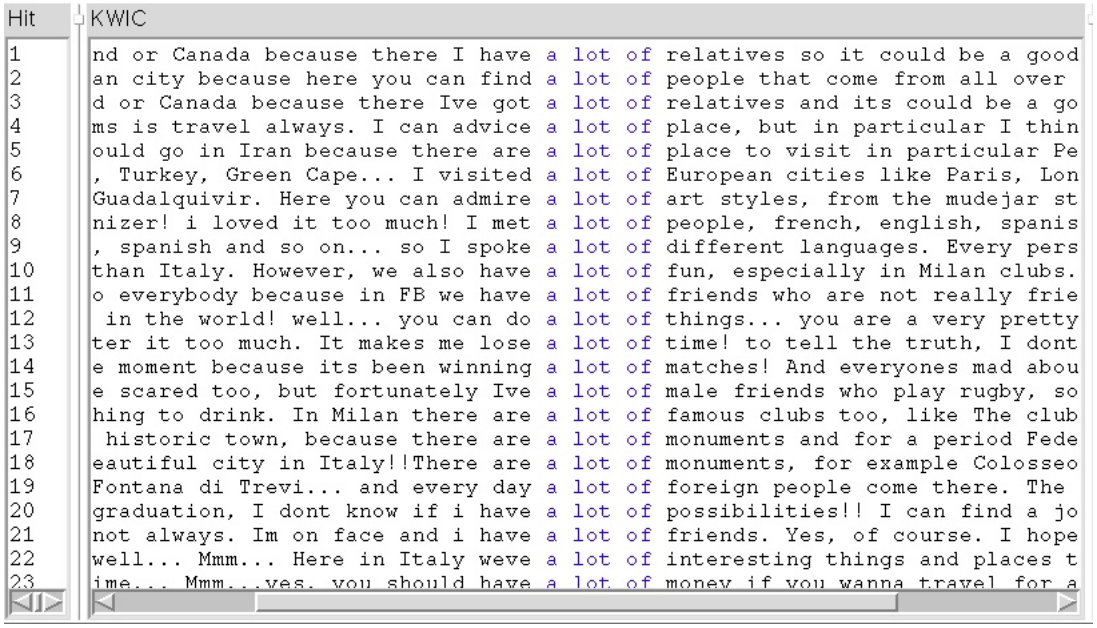
| Hit | KWIC |
|---|---|
| 1 | . Have you ever been in Italy? What about planning to go he |
| 2 | hat about planning to go here? There are so many places to |
| 3 | away the week stress. and you? where do you live? I lived u |
| 4 | study or do you already work? Im very interested in knowin |
| 5 | ountry! wouldnt it be amazing? Then, there is Australia. Bu |
| 6 | nglish and Russian! What else? Well, Im 22 and Im just a nc |
| 7 | s totally fun, dont you think? What about you? Tell me some |
| 8 | ont you think? What about you? Tell me something about you |
| 9 | lf, Im pretty curious! Really? My sister is in North Carol: |
| 10 | worth going, trust me! Really? What a pity! I love getting |
| 11 | m too one day! The best place? I dont know... its difficult |
| 12 | since it doesnt resemble you? Its just a suggestion, dont |
| 13 | Barcelona, Vienna and Prague? Ive been there and trust me: |
| 14 | What about Spanish or Chinese? I do like getting to know d: |
| 15 | r heard of Positano and Capri? See, lemons, caves and much |
| 16 | bout cous-cous, kebab, paella? My favourite dish is spaghet |
| 17 | to music. And what about you? Where do you live? I enjoy |
| 18 | prefer: mountains or seaside? If you like mountains I sugg |
| 19 | a little bit, dont you think? Im 21 years old and I study |
| 20 | t how do you really look like? Im curious! You were right, |
| 21 | Do you know what "Alpini" are? They are volunteers that du |
| 22 | dont you put a picture of you? I look like exactly my avat |

Most of the sequences produced by learners are part of independent clauses. The presence of dependent clauses is very low, the only exceptions being the fragment of a conditional clause, *if you want*, and two to-clause fragments, *to go to* and *to find a*. This fact indicates that the sentences produced by learners in the asynchronous chat are mainly simple sentences linked by means of coordination rather than subordination. This particular finding is not surprising, since the

syntax of computer-mediated English has been described as simplified (Herring, 2012). In CMC it is known that the interpersonal dimension takes precedence over the textual dimension (Danet and Herring, 2007) and communication is carried out by means of accumulation, the coordinator *and* and also of non-standard use of punctuation marks, suspension dots and exclamation marks in particular (Dresner and Herring, 2010).

The single most frequent 3-word sequence, *a lot of*, is a quantifier expression. Figure 4.10 shows concordances for this quantifier expression from the learner chats.

Figure 4.10: Concordance Lines for *a lot of* from LCC



| Hit | KWIC |
|---|---|
| 1 | nd or Canada because there I have a lot of relatives so it could be a good |
| 2 | an city because here you can find a lot of people that come from all over |
| 3 | d or Canada because there Ive got a lot of relatives and its could be a go |
| 4 | ms is travel always. I can advice a lot of place, but in particular I thin |
| 5 | ould go in Iran because there are a lot of place to visit in particular Pe |
| 6 | , Turkey, Green Cape... I visited a lot of European cities like Paris, Lon |
| 7 | Guadalquivir. Here you can admire a lot of art styles, from the mudejar st |
| 8 | nizer! i loved it too much! I met a lot of people, french, english, spanis |
| 9 | , spanish and so on... so I spoke a lot of different languages. Every pers |
| 10 | than Italy. However, we also have a lot of fun, especially in Milan clubs. |
| 11 | o everybody because in FB we have a lot of friends who are not really frie |
| 12 | in the world! well... you can do a lot of things... you are a very pretty |
| 13 | ter it too much. It makes me lose a lot of time! to tell the truth, I dont |
| 14 | e moment because its been winning a lot of matches! And everyones mad abou |
| 15 | e scared too, but fortunately Ive a lot of male friends who play rugby, so |
| 16 | hing to drink. In Milan there are a lot of famous clubs too, like The club |
| 17 | historic town, because there are a lot of monuments and for a period Fede |
| 18 | eautiful city in Italy!!There are a lot of monuments, for example Colosseo |
| 19 | Fontana di Trevi... and every day a lot of foreign people come there. The |
| 20 | graduation, I dont know if i have a lot of possibilities!! I can find a jo |
| 21 | not always. Im on face and i have a lot of friends. Yes, of course. I hope |
| 22 | well... Mmm... Here in Italy weve a lot of interesting things and places t |
| 23 | ime... Mmm...yes, you should have a lot of money if you wanna travel for a |

As can be seen in Figure 4.10, *a lot of* is used as a quantifier expression for a wide range of nouns, from *monuments* to *possibilities*. The constant repetition of *a lot of* provides corroborating evidence that learners employ a restricted number of sequences. Previous studies, such as De Cock (1998 and 2004), for example, have shown that learners repeatedly employ sequences that they consider safe. By holding onto expressions they can rely on, learners economise production effort, but their language appears to be much less varied as a result. Since the use of *a lot of* provides interesting insights into learner English, a cross-corpus comparison of *a lot of* will be carried out in section 5.3.2.

The description of sequence patterns above has confirmed that learner sequences are more similar to speech than writing. These findings, however, need to be complemented by a description of the functional features of learner sequences before generalisations can be made. The functional analysis is carried out in the next section.

### 4.2.2 Description of Functions

The present section presents a further classification of learners' recurrent sequences from LCC. The first twenty 3-word sequences, which were classified in terms of structures in the previous sections, are now classified in terms of functions. The functional description of the lexical bundles broadly follows Biber *et al.* (2004), who applied it to university teaching and writing. The same functional types are described and exemplified in Biber and Barbieri (2007). This classification was also used by De Cock (2004) in her study of recurrent sequences in the spoken learner corpus LINDSEI.

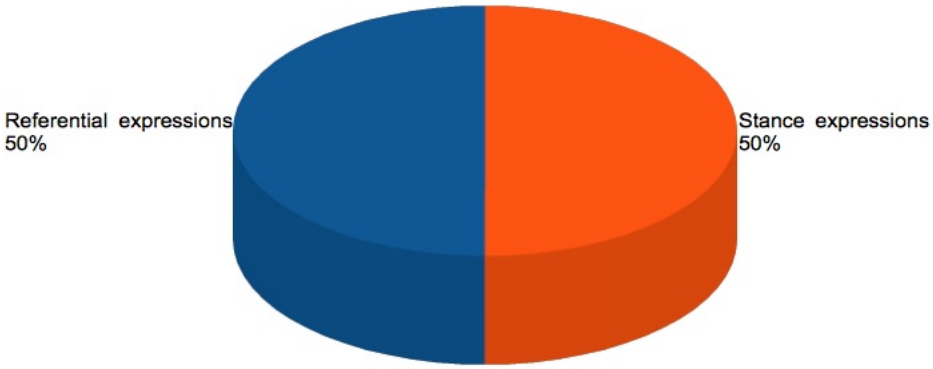
According to Biber and Barbieri (2007), recurrent sequences have three primary discourse functions: they are used as stance expressions, discourse organisers, and referential expressions. Stance expressions are used to frame another proposition with expressions of attitudes or assessments of certainty. They can be markers of attitudinal stance, markers of epistemic stance, responses, and markers of vagueness. Discourse organisers are used to reflect the relationships between previous and following parts of the discourse and they can be markers of speech or thought reporting, markers of contrast, markers of cause, and exemplifiers. Referential expressions are used to refer directly to physical or abstract entities or to the textual context. They include markers of time and place, quantifying sequences, and topic-dependent sequences. Table 4.3 shows the classification of LCC's 3-word sequences in terms of functions.

Figure 4.11 is a visual representation of the most frequent 3-word learner sequences by functions.

Table 4.3: 3-word Sequences in LCC: Functions

| Rank | Norm. Freq. | Sequence | Function |
|---|---|---|---|
| 1 | 205.9 | a lot of | Referential |
| 2 | 176.2 | i live in | Referential |
| 3 | 130 | i think that | Stance |
| 4 | 126.1 | would like to | Stance |
| 5 | 121.7 | i dont like | Stance |
| 6 | 120.7 | i would like | Stance |
| 7 | 113.4 | i dont have | Referential |
| 8 | 108.5 | with my friends | Referential |
| 9 | 107.1 | to go to | Referential |
| 10 | 87.1 | i dont know | Stance |
| 11 | 80.8 | id like to | Stance |
| 12 | 79.3 | find a job | Referential |
| 13 | 75 | i have to | Stance |
| 14 | 74.5 | i have a | Referential |
| 15 | 74.5 | live in a | Referential |
| 16 | 67.7 | like to go | Stance |
| 17 | 65.2 | go to the | Referential |
| 18 | 65.2 | i really like | Stance |
| 19 | 60.4 | to find a | Referential |
| 20 | 58.9 | if you want | Stance |

Figure 4.11: Functions of 3-word Sequences in LCC



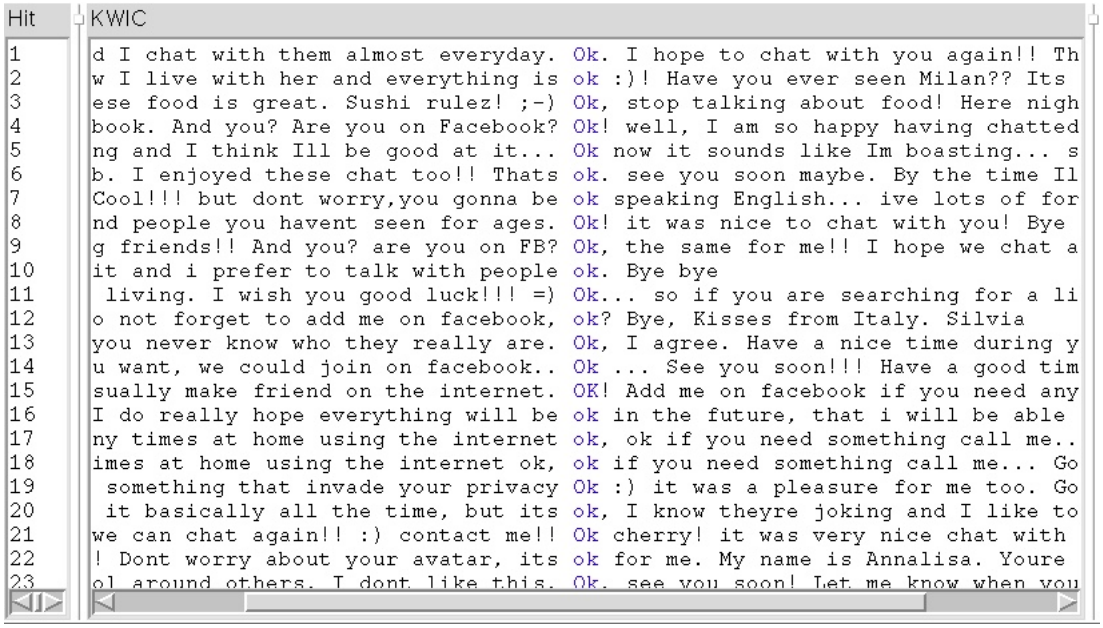

The most noticeable feature of the functional characteristics of the learners' recurrent sequences is that there is an equal presence of stance expressions and referential expressions (which include context-dependent sequences), while no discourse organising expression appears among the twenty most frequent sequences. Stance expressions, on the other hand, are particularly important in conversation, as they are used to express personal feelings, thoughts and desires, frequently in

the negative form. They also function as utterance launchers, presenting a personal stance relative to the information in the following clause. Biber *et al.* (1999) cite expressions such as *I don't know*, *I don't like*, *I don't think* and *I don't want*. The first two sequences can be found among LCC's most frequent, while *I don't think* is not present. Since stance expressions including the verb *think* are among the most frequently repeated in the chats, they are subjected to qualitative analysis and compared with *think*-sequences in the other two corpora in section 5.3.

Biber (2009) reports that native-speaker conversation is characterised by stance bundles. With regards to discourse organising bundles, Biber notes that they are frequent in conversation as they are used to indicate relationships between parts of the discourse. Biber's findings, however, are based on spoken university registers, which include spoken and written activities associated with university life, including classroom teaching. During lessons, professors repeatedly employ expressions for introducing, elaborating and clarifying topics. Using discourse organising sequences of this type would clearly be unnecessary in an interactive chat.

The absence of discourse organising sequences among the most frequent 3-word sequences indicates that learners organised their discourse using different expressions, including expressions shorter than 3-words. According to Biber *et al.* (1999:1046-7), in native-speaker conversation markers used to organise discourse are 'items loosely attached to the clause and connected with ongoing interaction'. Some of them are single word inserts such as *well*, *now* and *ok*, others are formulaic clausal forms such as *I mean* and *you know*. A quantitative analysis of concordance searches for *well* and *ok* evidenced that the learners employ the discourse markers *well* (351 concordances) and *ok/okay* (43 concordances) at clause boundaries. As can be seen in the concordance lines in Figure 4.12, many occurrences of *ok* are in fact discourse markers. Some learners even use *well* and *ok* one after the other, or in conjunction with suspension dots, emoticons, or interjections like *oh* and *er*.

Figure 4.12: Concordance Lines for *ok* from LCC



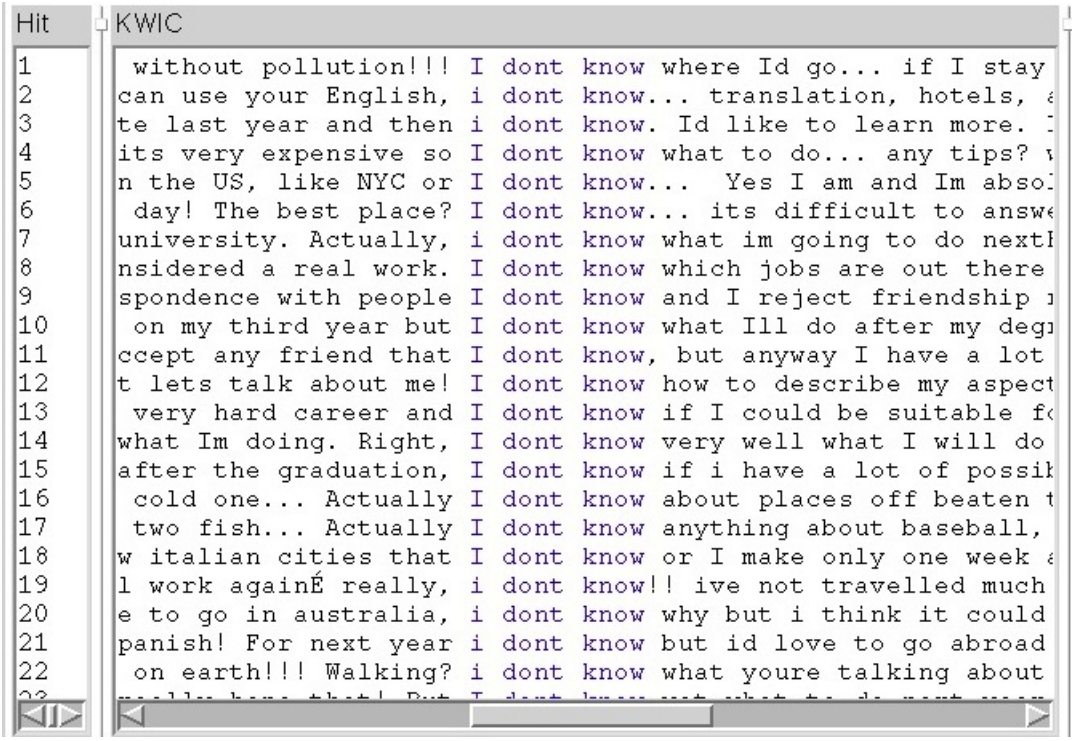
| Hit | KWIC |
|---|---|
| 1 | d I chat with them almost everyday. Ok. I hope to chat with you again!! Th |
| 2 | w I live with her and everything is ok :)! Have you ever seen Milan?? Its |
| 3 | ese food is great. Sushi rulez! ;-) Ok, stop talking about food! Here nigh |
| 4 | book. And you? Are you on Facebook? Ok! well, I am so happy having chatted |
| 5 | ng and I think Ill be good at it... Ok now it sounds like Im boasting... s |
| 6 | b. I enjoyed these chat too!! Thats ok. see you soon maybe. By the time Il |
| 7 | Cool!!! but dont worry,you gonna be ok speaking English... ive lots of for |
| 8 | nd people you havent seen for ages. Ok! it was nice to chat with you! Bye |
| 9 | g friends!! And you? are you on FB? Ok, the same for me!! I hope we chat a |
| 10 | it and i prefer to talk with people ok. Bye bye |
| 11 | living. I wish you good luck!!! =) Ok... so if you are searching for a li |
| 12 | o not forget to add me on facebook, ok? Bye, Kisses from Italy. Silvia |
| 13 | you never know who they really are. Ok, I agree. Have a nice time during y |
| 14 | u want, we could join on facebook.. Ok ... See you soon!!! Have a good tim |
| 15 | sually make friend on the internet. OK! Add me on facebook if you need any |
| 16 | I do really hope everything will be ok in the future, that i will be able |
| 17 | ny times at home using the internet ok, ok if you need something call me.. |
| 18 | imes at home using the internet ok, ok if you need something call me... Go |
| 19 | something that invade your privacy Ok :) it was a pleasure for me too. Go |
| 20 | it basically all the time, but its ok, I know theyre joking and I like to |
| 21 | we can chat again!! :) contact me!! Ok cherry! it was very nice chat with |
| 22 | ! Dont worry about your avatar, its ok for me. My name is Annalisa. Youre |
| 23 | ol around others. I dont like this. Ok. see you soon! Let me know when you |

Another concordance search shows that, in the whole corpus, *you know* is used as a discourse marker 58 times out of 130 concordances, and *I mean* 35 times out of 39 concordances. The *LGSWE* reports that the expression of stance *i don't know* is very common in speech because it can be used as an utterance launcher and as a discourse marker, chiefly to gain thinking time. The sample of concordance lines shown in Figure 4.13 reveals that in LCC *i don't know* is used in some instances to express stance (such as hits 1, 7 and 22), and in other instances as a marker of discourse (see hits 2, 5 and 6).

This preference for shorter expressions for organising discourse could be explained in two different ways: either learners choose shorter expressions to minimise typing effort, or they are using markers from speech because they consider them more suited to the electronic mode of communication.

Among the bundles that are very frequent in native-speaker conversation are bundles including the verbs *said* and *told* (Biber *et al.*, 1999). The LCC learners also mark their discourse with expressions including the verbs *said* and *told*. Whenever they find themselves repeating something from a previous turn, they use expressions such as *as i said*, *like i said*, *as you said*, and *as i told you*.

Figure 4.13: Concordance Lines for *i dont know* from LCC



| Hit | KWIC |
|---|---|
| 1 | without pollution!!! I dont know where Id go... if I stay |
| 2 | can use your English, i dont know... translation, hotels, |
| 3 | te last year and then i dont know. Id like to learn more. |
| 4 | its very expensive so I dont know what to do... any tips? |
| 5 | n the US, like NYC or I dont know... Yes I am and Im abso |
| 6 | day! The best place? I dont know... its difficult to answ |
| 7 | university. Actually, i dont know what im going to do next |
| 8 | nsidered a real work. I dont know which jobs are out there |
| 9 | spondence with people I dont know and I reject friendship |
| 10 | on my third year but I dont know what Ill do after my deg |
| 11 | ccept any friend that I dont know, but anyway I have a lot |
| 12 | t lets talk about me! I dont know how to describe my aspect |
| 13 | very hard career and I dont know if I could be suitable f |
| 14 | what Im doing. Right, I dont know very well what I will do |
| 15 | after the graduation, I dont know if i have a lot of possi |
| 16 | cold one... Actually I dont know about places off beaten t |
| 17 | two fish... Actually I dont know anything about baseball, |
| 18 | w italian cities that I dont know or I make only one week |
| 19 | l work againÉ really, i dont know!! ive not travelled much |
| 20 | e to go in australia, i dont know why but i think it could |
| 21 | panish! For next year i dont know but id love to go abroad |
| 22 | on earth!!! Walking? i dont know what youre talking about |

## 4.3 Conclusions

Summing up, the quantitative findings on recurrent sequences in a corpus of learner chats show that learner sequences mainly revolve around the pronoun *I* and active verb phrases. Among the most frequent sequences are verbs expressing stance, such as *think* and *know*. Overall, the frequency findings indicate that learner English in LCC is characterised by the repetitive use of a restricted number of sequences, some of which acquire a quasi-formulaic status. In terms of register, the findings show that learner CMC English shares some features of native-speaker conversation. Since CMC is generally considered closer to speech than writing, this might be considered evidence that learners produce sequences that are consistent with the mode of communication they are employing.

In terms of structure, learner sequences most frequently include personal pronouns and active verb phrases, while there are very few noun and prepositional phrases. The structural analysis, on the other hand, shows that in the chats the learners mainly produced simple sentences linked by means of coordination rather

than subordination. Since no question fragments appear in the top twenty recurrent sequences, a concordance search was carried out to ascertain whether the learners asked any questions. The concordance data show that the learners who took part in the chat did ask questions using a wide range of structures, many of which are non-standard structures, typical of speech. Overall, the structures of learner sequences can be described as more speech-like than writing-like.

The classification of recurrent sequences by function showed that the twenty most frequent sequences of words in the corpus are equally divided into stance and referential expressions and that there are no discourse organising sequences among the twenty 3-word sequences with the highest occurrences. A qualitative investigation of the corpus revealed that learners use shorter expressions to organise and link their thoughts.

The next step of the present research study is the comparison of the sequences retrieved from LCC with those retrieved from ICLE_IT, a corpus of learner essays produced by Italian learners, and LINDSEI_IT, a small corpus of learner interviews produced by Italian learners. The aim is to discover whether, and to what extent, learner English sequences differ across the three modes of communication.

# Chapter Five

# Comparison with ICLE and LINDSEI

In the present chapter, the data from LCC is compared to two other corpora of learner English: ICLE-IT, ICLE's Italian subcorpus, a corpus of learner essays, and LINDSEI-IT, LINDSEI's Italian subcorpus, a collection of spoken interviews. By means of corpus comparison, the study addresses register differences in learner English and seeks an answer to the final research questions on learners' recurrent sequences across different modes of communication. In the analysis of LCC an influence of the medium on the language produced could be observed by looking at the frequency data. ICLE_IT's and LINDSEI_IT's frequency analyses are expected to show the same phenomenon.

As reviewed in Chapter Two, over the past decade several studies have compared learner English with native-speaker English by means of comparable corpora. These comparisons have provided a description of learner language in terms of underuse, overuse and misuse of specific lexis, structures and sequences of words. The aim of the present chapter, however, is different. It is the description of recurrent sequences in learner English across different registers without referring to native-speaker models. In particular, it is an analysis of learner language across registers within the framework of analysis expounded by Biber *et al.*

(1999) in the LGSWE and in his successive studies of lexical bundles in native-speaker English. A cross-register comparison is a very useful tool, in that, by comparing learner language with itself, it can give us an insight into language processing in different contexts. In addition, it will be fairer than comparisons with the native-speaker model, which, after all, modern scholars largely consider unattainable.

After the description of the most frequent learner sequences in terms of structures and functions, this section compares findings regarding LCC with recurrent learner sequences recorded in ICLE_IT, a corpus of essays and exam papers, and LINDSEI_IT, a corpus of oral interviews. The comparison is restricted to texts and interviews by learners with the same mother tongue background as those recorded in LCC, namely Italian. This decision is justified by previous findings on learner English: most studies have shown that mother tongue background has a great influence on learner language, not only because it accounts for interference of the native language, but also because differing learning contexts result in different learner Englishes, which are hardly comparable. As expounded in Section 3.1, learners' experiences with the English language in different countries vary considerably and, for the present study, it was considered essential for the learner samples to share the same variables in terms of context.

Section 5.3 is a qualitative cross-corpus comparison of two specific sequences: the most frequent sequences including the verb *think* and the quantifier *a lot of*. These sequences are particularly interesting because the quantitative analysis shows they appear among the most frequent in all the corpora. In particular, it is shown how the sequences are employed differently in the three corpora.

## 5.1 Quantitative Comparison

Corpus data for the three corpora compared in the present study are summarised in Table 5.1.

In terms of corpus size, the number of texts and average length LCC and

Table 5.1: LCC, ICLE_IT and LINDSEI_IT Corpus Data

| ***CORPUS*** | ***LCC*** | ***ICLE_IT*** | ***LINDSEI_IT*** |
|---|---|---|---|
| Corpus size (words) | 205,451 | 227,085 | 59,573 |
| Learner Texts | 253 chats | 392 essays | 50 interviews |
| Average Length | 812 | 580 | 1,191 |

ICLE_IT are very similar, even though the total number of essays included exceeds that of the chats and the average essay length is inferior to the average chat length. This latter difference could be explained by the fact that essays normally have a word limit, while, as explained in Section 3.3.3, the chat task had no set time nor a minimum or maximum length in terms of words.

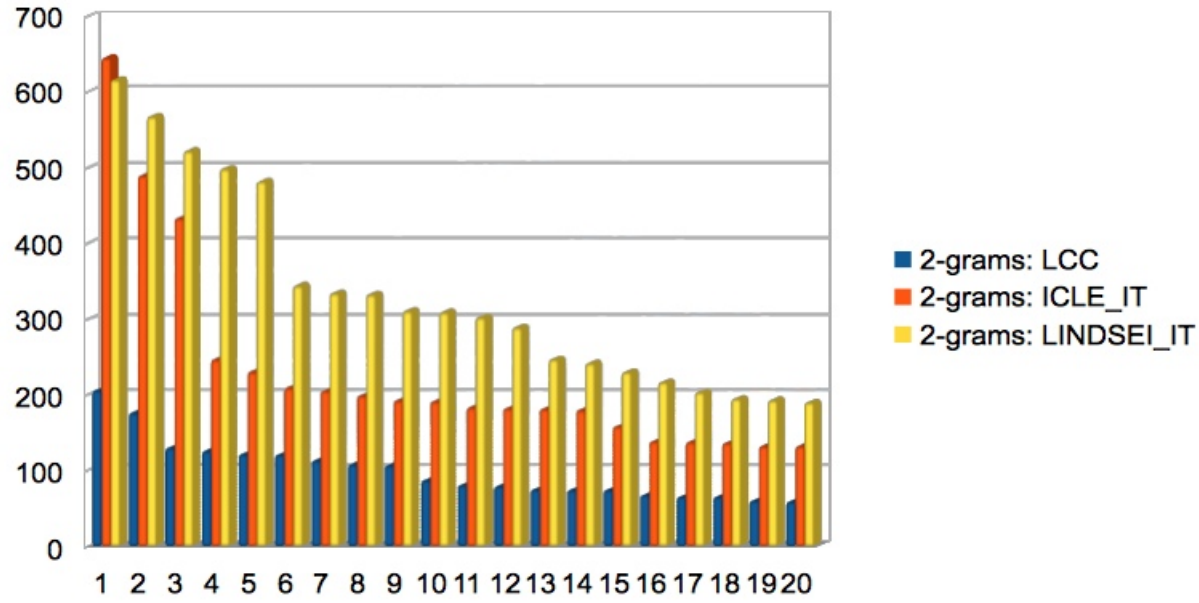
LINDSEI_IT, on the other hand, is considerably smaller than LCC and ICLE_IT, both in terms of a word count and the number of learner texts. This might be due to the fact that even though the length of the interview can be a set criterion, the amount of speech produced depends on a great number of variables, such as the number of interruptions, the length of pauses and the speech rate of each learner (see 3.2.2).

In order to have comparable data from all three corpora, N-grams were extracted from ICLE_IT and LINDSEI_IT with the same settings as were used for LCC. Due to the inferior size of LINDSEI_IT, the cut-off point for sequence extraction was set at 10, rather than 20. After extraction, all the frequencies were normalised to 100,000 words. The graphs in Figures 5.1, 5.2 and 5.3 present the normalised frequency data for 2- to 4-word sequences extracted from the three corpora (in the figures, the word n-gram is employed instead of n-sequence to save space).

As can be seen in Figure 5.1, there are considerably higher frequencies for 2-word sequences in ICLE_IT and LINDSEI_IT than in LCC. With regard to 2-grams, therefore, the quantitative data indicates that LCC texts tend to repeat the same sequences much less frequently than the texts included in the other two corpora. A closer analysis of the 2-word sequences explains the results.

In LINDSEI_IT, there are several of repetitions such as *the the*, *to to*, *i i* and *a*

Figure 5.1: 2-grams in LCC, ICLE_IT and LINDSEI_IT




*a*, which are typical dysfluencies of speech, but which do not represent a sequence in terms of the present study. The same can be said of sequences containing fillers like *eh*, *mm*, *er*, and so on. This makes 2-word sequences automatically extracted from LINDSEI_IT unsuitable for analysis, as they do not represent sequences of words in terms of the present research study. This case clearly exemplifies how automatically extracted quantitative data needs to be interpreted in the light of qualitative analysis before it can yield meaningful results.

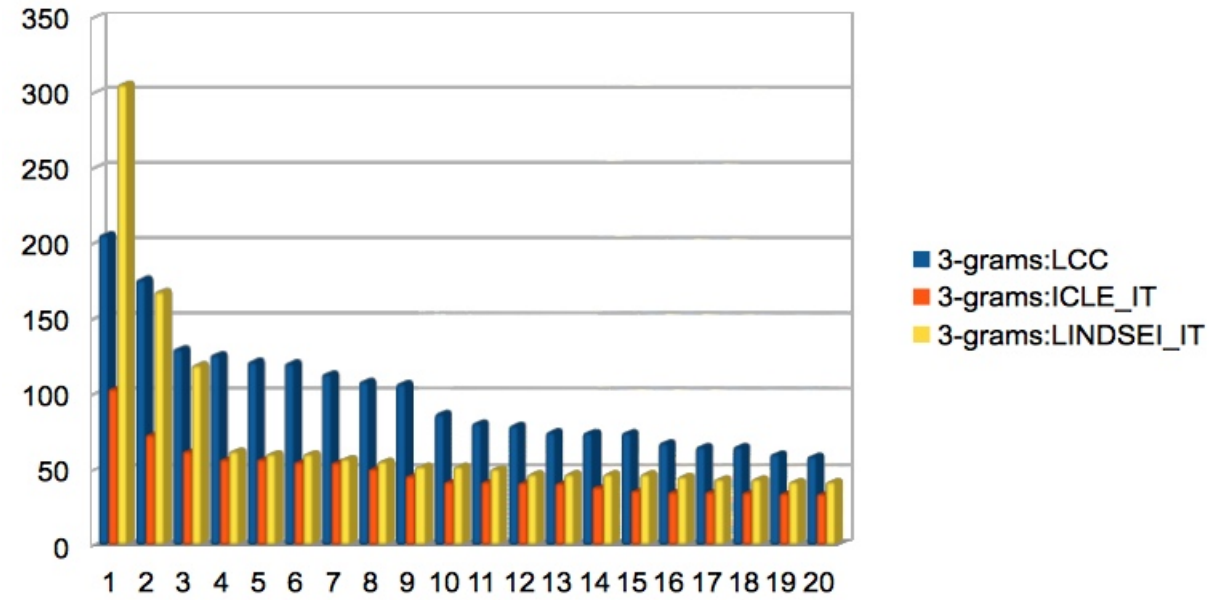
If we exclude dysfluencies, the most frequent 2-word sequences in LINDSEI_IT are *in the*, *i think*, *i don't*, *and so* and *don't know*. With the exception of the prepositional phrase fragment, these sequences are included in the most frequent bundles of conversation. As for *in the*, it can be explained by looking at 3- and 4-word sequences, since among the most repeated ones, in fact, are *in the first* and *in the first picture*. Clearly this sequence was used by learners at the beginning of the picture description task.

As for ICLE_IT, the two most frequent 2-word sequences, *of the* and *in the*, are the beginning of a prepositional phrase. There are other frequent 2-grams which have the same structure: *to the*, *on the* and *with the*. In Biber *et al.* (1999)

the noun phrase and *of*-phrase fragments and prepositional phrase fragments in general are reported as being the distinctive structural type of academic prose. The sequence *of the*, for example, can be used to form bundles such as *the beginning of the*, *part of the*, *one of the most*, and so on, which are frequently used in academic writing. As a matter of fact, *one of the*, *is one of the* and *one of the most* can be found among the top twenty 3- and 4-word sequences of ICLE_IT.

The most recurrent verb among the 2-word sequences extracted from ICLE_IT is *be*, in the present forms and in the base and to-infinitive forms. The only other verbs are *have* and the modal *should*. Not surprisingly for written English in an academic context, there are no contracted forms. The only two nouns among the most repeated sequences are *insemination* and *child*. As ICLE_IT is a corpus of argumentative essays on predetermined topics, this explains the high frequency of specific content words. Overall, 2-word sequences in ICLE_IT show features that are found in native-speaker formal writing.

Figure 5.2: 3-grams in LCC, ICLE_IT and LINDSEI_IT




As can be seen in Figure 5.2, the frequency picture changes with 3-word sequences: only the most frequent one extracted from LINDSEI_IT, i.e. *i don't know*, is used more repeatedly than 3-word sequences in LCC. The others present

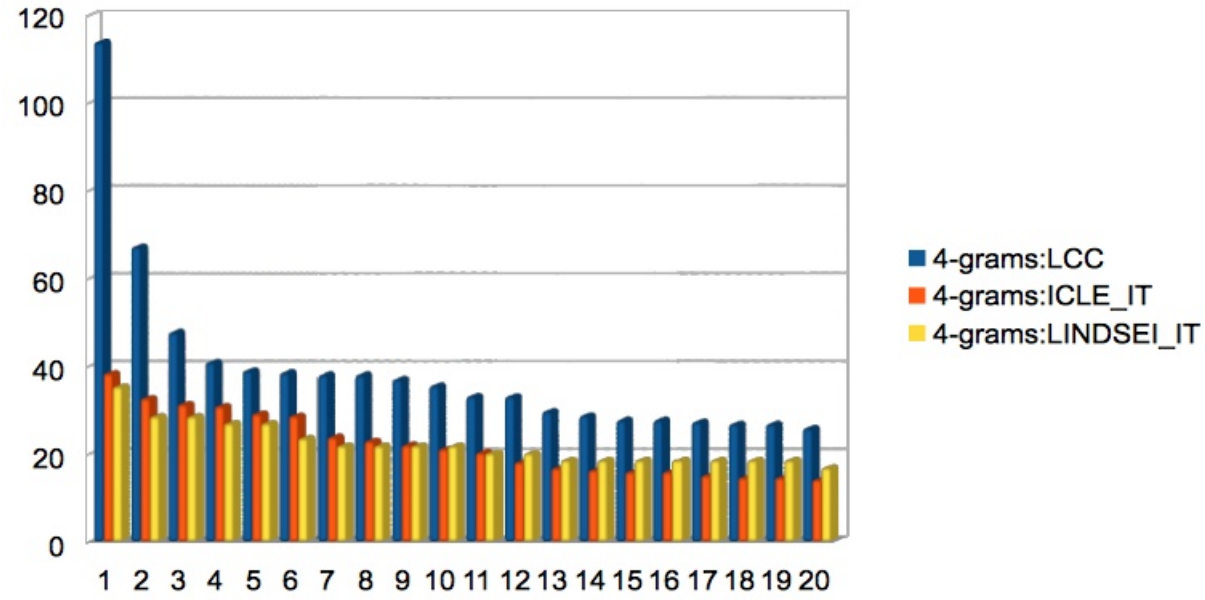
frequencies which are equal or inferior to the frequencies in LCC. As for ICLE_IT, 3-word sequences are consistently used less frequently than in LCC.

With regard to 3-word sequences, the most frequently repeated combination across the three corpora investigated is *i dont know*, which appears in LINDSEI_IT with a normalised frequency 182 occurrences per 100,000 words. As reported in Biber *et al.* (1999), in conversation this sequence is often employed as a discourse marker and, as we have seen in Section 4.2.2, it also appears among the most frequent sequences in LCC. Among the verbs, *think* and *know* are at the top of the list, accompanied by *was*. The *LGSWE* reports that personal verbs reporting feelings and thoughts are very commonly used to begin utterances in spoken English. There are two sequences with the contracted form *don't* and linking words such as *and*, *so* and *because*, all of which are found in recurrent sequences from conversation.

Other frequent combinations in LINDSEI_IT are *i think that* and *a lot of*. Both these sequences are found in LCC and in ICLE_IT and will be the subject of qualitative comparison in section 5.3. In quantitative terms, however, with the exception of *i dont know*, LCC has more 3-word and 4-word sequences than the other two corpora. In order to analyse the three corpora following the same methodology, 3-word sequences from ICLE_IT and LINDSEI_IT were also analysed in terms of structures and functions, therefore they are discussed in greater detail in Section 5.2.

Figure 5.3 presents a graph showing 4-word sequences from the three corpora. As can be seen, sequences of four words are more frequently repeated in LCC than in the other two corpora. 4-word sequences from ICLE_IT are context-bound, and include a large number of content words such as *artificial insemination*, *child*, *children*, *single women* and *baby*. They are also characterised by a high frequency of the verbs *be allowed to* and *should*, which appear in six out of twenty recurrent sequences. As will be clear from the analysis of 5- and 6-grams from ICLE, the repetition of these verbs is connected to a specific argumentative essay title. Other

Figure 5.3: 4-grams in LCC, ICLE_IT and LINDSEI_IT

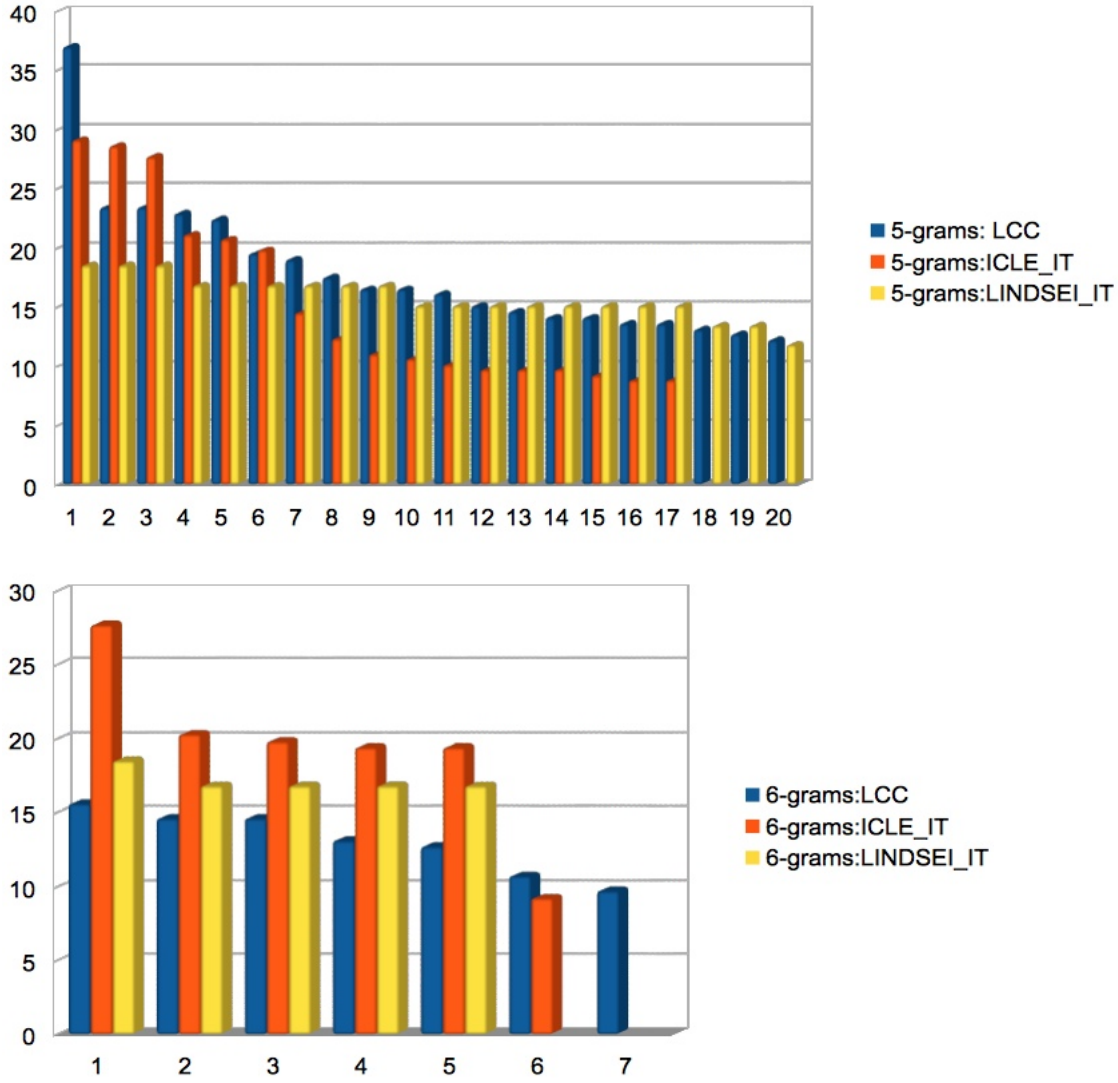



interesting sequences in ICLE_IT are *on the other hand*, *as a matter of*, *a matter of fact* and *at the same time*, which are used in the essays as discourse organisers. Overall, learners in written essays use discourse organisers which are more typical of the register. In LCC, instead, they tended to organise their thoughts using shorter discourse markers.

In LINDSEI_IT, the most frequent 4-word sequences are more typical of speech. Similarly to the findings regarding 2-word sequences, the automatic analysis retrieved four sequences including repetitions and hesitation markers. Among the most frequently repeated 4-grams are *i don't know i* and *i would like to*. Indeed, six out of twenty 4-grams extracted from LINDSEI_IT include the expression *don't know*. As we mentioned before, combinations of *don't know* with the pronoun *I* are a staple of conversation as they are used for hedging and as discourse markers.

The graphs in Figure 5.4 show the normalised frequency data (per 100,000 words) for the 5- and 6-word sequences in the three corpora. ICLE_IT has higher frequencies, but of a smaller number of sequences: in fact, only seventeen sequences have raw frequencies higher than 20. Examples from ICLE_IT are *should not be allowed to*, *allowed to have artificial insemination* and *be allowed*

Figure 5.4: 5- and 6-grams in ICLE_IT and LINDSEI_IT




*to have artificial*. Clearly, they are referential sequences that can be traced back to the essay title. Only six 6-word sequences with raw frequencies higher than 20 were extracted from ICLE_IT. Overall, they exhibit very similar features to 5-word sequences, and are rephrasings of the title of an essays, examples being *be allowed to have artificial insemination* and *women should not be allowed to*.

Similarly, recurrent 5- and 6-word combinations in LINDSEI_IT have lower frequencies than LCC and ICLE_IT. Examples from LINDSEI_IT are *a country you have visited*, *a country you have visited which, describe your visit and say* and *you found the country particularly impressive*. In these sequences, the speaker interviewed is reading or repeating the interview topic. In LINDSEI_IT, 5- and

6-word sequences are not produced by learners as novel expressions, but are rather verbatim repetitions of the task. Longer sequences are often dependent on context and the result of local repetition.

A similar picture was uncovered for LCC. A considerable number of 5- and 6-word sequences are formulae and salutations, such as, *hi cherry nice to meet you* and *hi cherry my name is*. Others echoed the turn the learner has just read, as, *make friends over the internet*, or *the best place ive ever been*.

To sum up, the quantitative findings indicate that recurrent sequences in the learner English corpora investigated are generally appropriate to the register and the means of communication being used: learner writing features a higher number of prepositional and noun phrases, while learner speech makes extensive use of the personal pronoun *I* and discourse markers. In addition, the corpora provide evidence that a large number of 5- and 6-word sequences are not actually assembled by the learners, because they are repetitions of rubrics, such as essay titles and interview questions. As such, they are not very interesting for the purposes of a study of recurrent sequences in learner English, and the qualitative comparisons in the present study thus focus on 3-word combinations.

## 5.2 Patterns and Functions

As one of the aims of the present study is to find out whether learners are able to adapt their language to the mode of communication they adopt, in the present section the most frequent 3-word sequences extracted from LCC, ICLE_IT and LINDSEI_IT are compared in terms of patterns and functions.

### 5.2.1 ICLE_IT

Table 5.2 lists the twenty most frequent 3-word sequences automatically extracted from ICLE_IT classified into structural types. Frequencies are normalised per 100,000 words. This analysis provides interesting data for comparison with LCC.

Table 5.2: 3-word Sequences: Patterns and Functions in ICLE-IT

| N.F. | Sequence | Structure | Function |
|---|---|---|---|
| 103.5 | in order to | To-clause fragm. (subordin.) | Referential |
| 73.1 | to have a | To-clause | Referential |
| 62.1 | it is not | Pronoun + VP | Stance |
| 56.8 | one of the | Noun Phr. + of phrase fragm. | Referential |
| 56.8 | the fact that | Noun Phr. + other post-mod. | Disc. org. |
| 55 | should not be | Verb Phrase (modal) | Stance |
| 54.6 | be allowed to | Verb Phrase (passive) | Stance |
| 50.6 | i think that | Pronoun + VP | Stance |
| 45.8 | a lot of | Noun Phr. + of phrase fragm. | Referential |
| 41.8 | for their children | Prepositional Phrase | Referential |
| 41.8 | have artificial insemination | (Verb Phrase) + Noun phr. | Referential |
| 41.4 | in my opinion | Prepositional Phrase | Stance |
| 41 | point of view | Noun Phr. + of phrase fragm. | Referential |
| 38.3 | to have artificial | To-clause | Referential |
| 36.1 | on the other | Prepositional Phrase | Disc. org. |
| 35.2 | a single woman | Noun Phrase | Referential |
| 35.2 | first of all | Adverbial + of phrase | Disc. org. |
| 34.8 | allowed to have | Verb Phrase (passive) | Stance |
| 34.3 | single women should | Noun Phrase + VP fragm. | Stance |
| 33.9 | the number of | Noun Phr. + of phrase fragm. | Referential |

In terms of structure, the main differences can be summarised as follows:

1. Only two sequences feature personal pronouns, namely *it* and *i*. Sequences in LCC, on the contrary, consist predominantly of 1st person pronoun + verb phrase sequences.

2. Noun phrases and prepositional phrases are predominant (50%), while they make up only 10% of the recurrent word sequences in LCC.

3. The passive verb form *be allowed to* is among the most repeated structures, while there are no passive verb forms among the most frequent sequences in LCC.

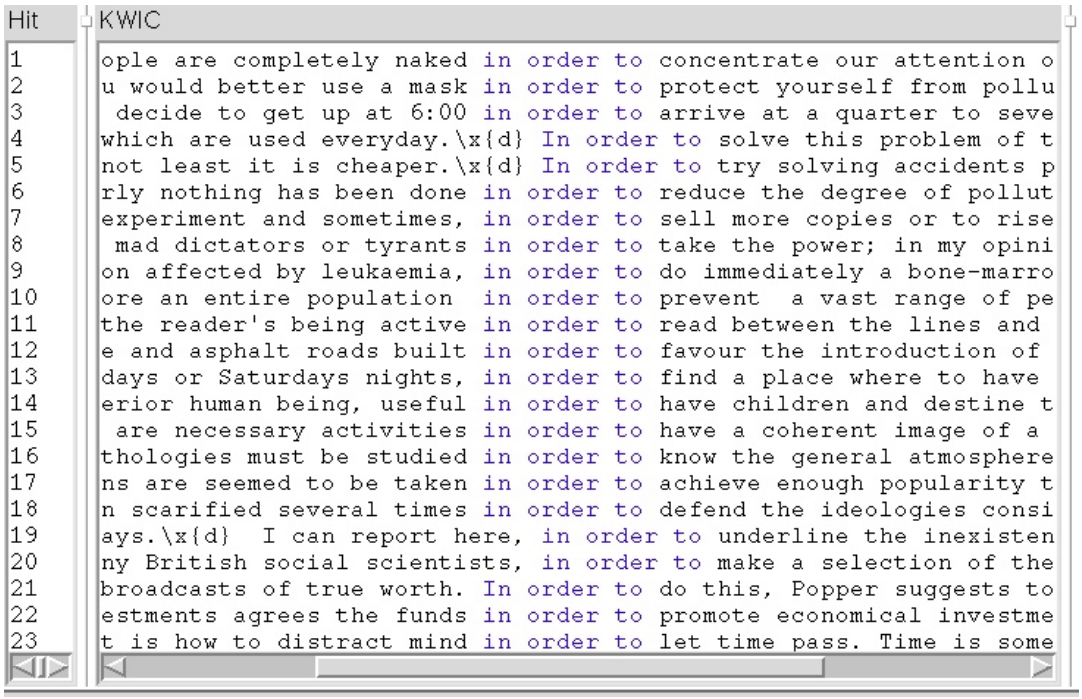
4. Sequences are structurally incomplete, whereas in LCC many 3-word combinations include subject, verb phrase and, in some cases also the following noun phrase.

Overall, the structures of the sequences in ICLE_IT are consistent with the written register. This is particularly true of the 3-word sequence *in order to*. Biber *et al.* (1999) define it a subordinator of purpose and corpus-based findings from LGSWE indicate that it is commonly employed in academic English either in initial or in mid position. According to Biber *et al.* (1999:839), expository written registers tend to use more subordinators followed by a non-finite clause because they confer precision to the communication by making the relationships between the clauses explicit. Other combinations used with the same purpose is *so as to*, whose occurrences in ICLE_IT are fewer, seventeen in total. In conversation, instead, purpose clauses normally begin with the infinitive marker *to*.

Since *in order to* is the most frequently repeated 3-word sequence in ICLE_IT, it deserves attention. On the one hand, the repeated use of this subordinator indicates that learner English in writing has a fairly complex syntax, and subordinate clauses are often introduced by a sequence which is typically used in writing. On the other hand, however, the overly repetitive use of *in order to* has the effect of conferring to it a quasi-formulaic status. It is an expression that learners seem to have memorised and use repeatedly whenever they have the occasion to do so. Further proof that this might be the case is that the expression can be found up to four to six times in the same essay. In other words, some learners seem to rely exceedingly on this preassembled phrase. This finding provides further evidence of the repetitiveness of learner language highlighted in previous studies.

Figure 5.5 shows a random sample of 23 out of the 235 concordances for *in order to* from ICLE_IT (corresponding to a normalised frequency of 103.5). The data show a preference for mid rather than initial position and in some sentences, such as *to get up at 6:00 in order to arrive at a quarter to*, and *completely naked in order to concentrate our attention*, it could be substituted by the infinitive marker *to*, and the sentence would be just as clear.

Figure 5.5: Sample Concordances for *in order to* from ICLE_IT



| Hit | KWIC |
|---|---|
| 1 | ople are completely naked in order to concentrate our attention o |
| 2 | u would better use a mask in order to protect yourself from pollu |
| 3 | decide to get up at 6:00 in order to arrive at a quarter to seve |
| 4 | which are used everyday.\x{d} In order to solve this problem of t |
| 5 | not least it is cheaper.\x{d} In order to try solving accidents p |
| 6 | rly nothing has been done in order to reduce the degree of pollut |
| 7 | experiment and sometimes, in order to sell more copies or to rise |
| 8 | mad dictators or tyrants in order to take the power; in my opini |
| 9 | on affected by leukaemia, in order to do immediately a bone-marro |
| 10 | ore an entire population in order to prevent a vast range of pe |
| 11 | the reader's being active in order to read between the lines and |
| 12 | e and asphalt roads built in order to favour the introduction of |
| 13 | days or Saturdays nights, in order to find a place where to have |
| 14 | erior human being, useful in order to have children and destine t |
| 15 | are necessary activities in order to have a coherent image of a |
| 16 | thologies must be studied in order to know the general atmosphere |
| 17 | ns are seemed to be taken in order to achieve enough popularity t |
| 18 | n scarified several times in order to defend the ideologies consi |
| 19 | ays.\x{d} I can report here, in order to underline the inexisten |
| 20 | ny British social scientists, in order to make a selection of the |
| 21 | broadcasts of true worth. In order to do this, Popper suggests to |
| 22 | estments agrees the funds in order to promote economical investme |
| 23 | t is how to distract mind in order to let time pass. Time is some |

Interestingly, the sequence *in order to* was also found in LCC, with a normalised frequency of 42.8, and in LINDSEI_IT, with a normalised frequency of 47. Frequencies are decidedly lower in these two corpora, especially in LCC, which has the lowest. However, while in written production it has the highest frequency and, in most cases, it is used correctly, in speech and in chats it is rather out of place in terms of the mode of communication. Sample concordance lines from LCC are:

CONC_19 *the Trevi Fountain (where several tourists go in order to throw a coin which is, according to a local legend*

CONC_39 *from other parts of the world, just like you, in order to know new things, new ideas and new points of view*

CONC_57 *and some of them decided to remain here in order to have a go. But i would also suggest to avoid*

Sample concordance lines from LINDSEI_IT are:

CONC_17 *go to Turin (eh) when I have some time (er) in order to see my boyfriend otherwise we cannot see each other*

CONC_18 *when I (erm) especially when I am abroad in order to see all those (eh) plays which I have studied*

CONC_20 *to know her father she wanted to know hi= him in order to be sure that the baby would have been clever and*

The frequency data from the three corpora suggest that this three-word phrase is one of those 'islands of reliability' Conklin and Schmitt (2008:76) that some learners cling on to when they use English, regardless of the register.

In terms of functions, three discourse organising sequences were extracted automatically from ICLE_IT, namely *the fact that*, *on the other* (followed by *hand*), and *first of all*. This finding indicates that learners tried to organise their essays following the conventions of written English. Conversely, as we saw in Chapter Four, the learners recorded in LCC did not use 3-word discourse organisers.

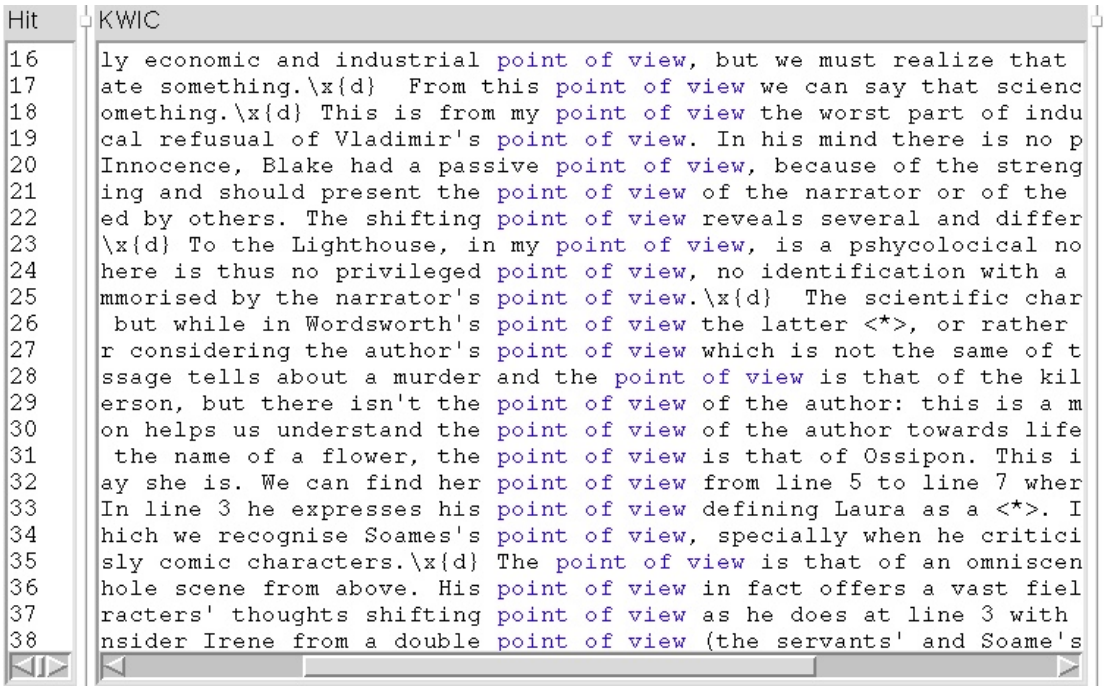
On the other hand, stance expressions are not as prominent in ICLE_IT as they were in LCC. In ICLE_IT, learners express stance by means of the sequences *i think that*, *in my opinion*, and through the use of modals, as in the sequence *should not be*. In addition, six occurrences of the sequence *point of view*, combined with *my*, are employed to express stance, as can be seen in sample concordances 18 and 23 in Figure 5.6.

It could be argued that when they express themselves in writing, the learners have preassembled sequences that they employ for the specific functions, such as discourse organising or expressing stance. As a result, however, their English, though appropriate in terms of register, is more repetitive.

### 5.2.2 LINDSEI_IT

For LINDSEI_IT the analysis of structures and functions was carried out on ten recurrent sequences of three words, that is only the most recurrent sequences that did not contain markers of speech dysfluencies (such as the repetitions *in in the* and the fillers *and eh the*). The classified sequences are presented in Table 5.3.

Figure 5.6: Concordances for *point of view* from ICLE_IT



| Hit | KWIC |
|---|---|
| 16 | ly economic and industrial point of view, but we must realize that |
| 17 | ate something.\x{d} From this point of view we can say that scienc |
| 18 | omething.\x{d} This is from my point of view the worst part of indu |
| 19 | cal refusual of Vladimir's point of view. In his mind there is no p |
| 20 | Innocence, Blake had a passive point of view, because of the streng |
| 21 | ing and should present the point of view of the narrator or of the |
| 22 | ed by others. The shifting point of view reveals several and differ |
| 23 | \x{d} To the Lighthouse, in my point of view, is a pshycolocical no |
| 24 | here is thus no privileged point of view, no identification with a |
| 25 | mmorised by the narrator's point of view.\x{d} The scientific char |
| 26 | but while in Wordsworth's point of view the latter <*>, or rather |
| 27 | r considering the author's point of view which is not the same of t |
| 28 | ssage tells about a murder and the point of view is that of the kil |
| 29 | erson, but there isn't the point of view of the author: this is a m |
| 30 | on helps us understand the point of view of the author towards life |
| 31 | the name of a flower, the point of view is that of Ossipon. This i |
| 32 | ay she is. We can find her point of view from line 5 to line 7 wher |
| 33 | In line 3 he expresses his point of view defining Laura as a <*>. I |
| 34 | hich we recognise Soames's point of view, specially when he critici |
| 35 | sly comic characters.\x{d} The point of view is that of an omniscen |
| 36 | hole scene from above. His point of view in fact offers a vast fiel |
| 37 | racters' thoughts shifting point of view as he does at line 3 with |
| 38 | nsider Irene from a double point of view (the servants' and Soame's |

In terms of structure, the following features can be seen in the recurrent features extracted from LINDSEI_IT:

1. Most of the 3-word sequences are structurally incomplete and they are not context-dependent.

2. There is a prevalence of the first and third person pronoun + verb phrase structure.

3. Noun and prepositional phrases are rare.

Table 5.3: 3-word Sequences Patterns and Functions in LINDSEI_IT

| **Norm. Freq.** | **Sequence** | **Structure** | **Function** |
|---|---|---|---|
| 305.5 | i dont know | 1st person pronoun + VP | Stance |
| 167.9 | i think that | 1st person pronoun + VP | Stance |
| 119.2 | a lot of | Noun Phrase | Referential |
| 60.4 | there is a | Noun Phrase + be | Referential |
| 57.1 | at the end | Prepositional phrase | Discourse org. |
| 55.4 | to go to | Verb Phrase | Referential |
| 52 | in the first | Prepositional phrase | Discourse org. |
| 52 | it was a | 3rd person pronoun + VP | Referential |
| 47 | i went to | 1st person pronoun + VP | Referential |
| 47 | in order to | To-clause fragm. (subordin.) | Referential |

In the other two corpora, recurrent 3-word sequences were mostly structurally incomplete, but they were considerably more context-dependent. As for personal pronouns, LINDSEI_IT is more similar to LCC than to ICLE_IT. Moreover, LINDSEI_IT shows a markedly lower presence of noun and prepositional phrases, especially in comparison with ICLE_IT.

In terms of verbs, *know* and *think* occur in the present tense, while *be* and *go* occur in the non-finite, present and past forms. LINDSEI_IT is the only corpus which shows a mix of past and present tenses. This is connected to the task the learners had to perform, which was to recount a film or play they had seen. No past forms appear among the most recurrent sequences in LCC, as the chats deal mostly with learners' present life and their plans for the future. In ICLE_IT, on the other hand, sequences reflect a greater use of dependent and non-finite clauses, and no past tense was featured among the most frequently repeated ones.

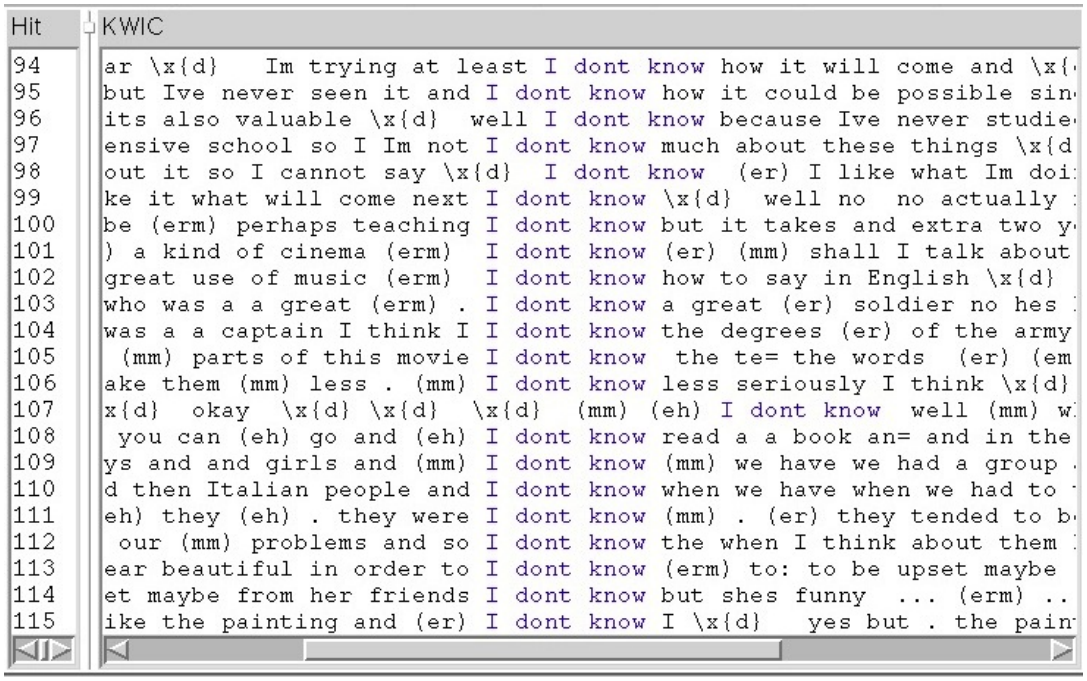
In terms of functions, the two sequences expressing stance, *i dont know* and *i think that*, are by far the most frequently repeated in the corpus. They are followed by referential sequences (including the quantifying sequence *a lot of*, which will be analysed in detail in Section 5.3.2) and there are two discourse organising sequences: *at the end* and *in the first*.

The discourse organisers employed by learners are closely connected to the two interview tasks: the narration of the film or play and the set picture description. Learners used the discourse organiser *at the end* to relate the story of the film and the expression *in the first* to describe the first picture of the set. It is worth noting once more that learners tend to use prefabricated sequences for discourse organising functions. This is typical of speech, as prefabricated sequences reduce processing effort and increase thinking time.

In this respect, an interesting sequence is *i dont know*, which also appeared among the most repeated sequences in LCC. In LINDSEI_IT the sequence is used as a discourse marker, a linguistic device to gain processing time, as can be seen

in the sample concordance lines in Figure 5.7. The same function is shared by pauses, repeats and hesitation items, used by learners to ease planning pressure, as noted by De Cock (2004).

Figure 5.7: Sample Concordances for *i dont know* from LINDSEI_IT



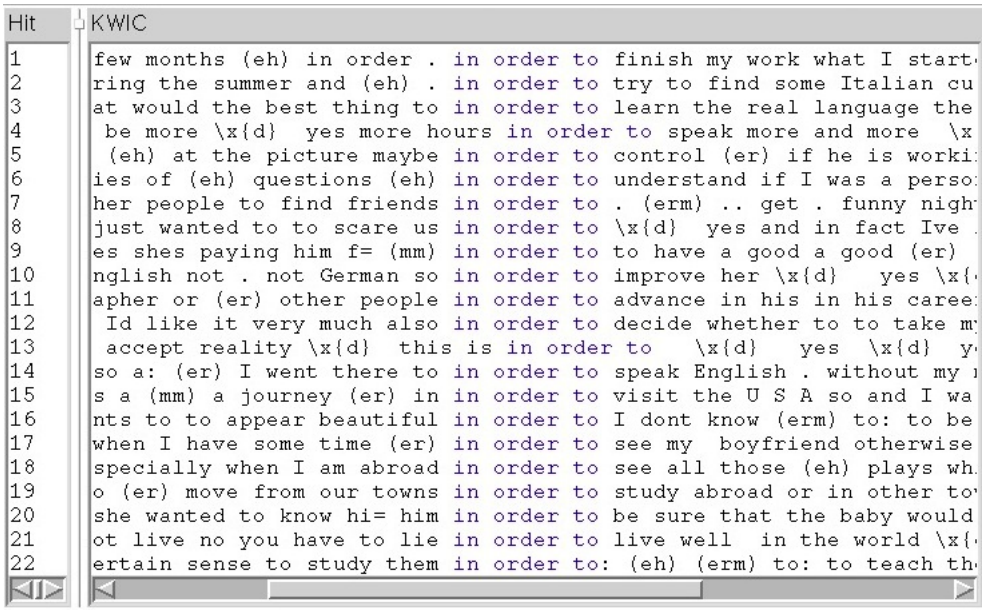
| Hit | KWIC |
|---|---|
| 94 | ar \x{d} Im trying at least I dont know how it will come and \x{ |
| 95 | but Ive never seen it and I dont know how it could be possible sin |
| 96 | its also valuable \x{d} well I dont know because Ive never studie |
| 97 | ensive school so I Im not I dont know much about these things \x{d |
| 98 | out it so I cannot say \x{d} I dont know (er) I like what Im doi |
| 99 | ke it what will come next I dont know \x{d} well no no actually |
| 100 | be (erm) perhaps teaching I dont know but it takes and extra two y |
| 101 | ) a kind of cinema (erm) I dont know (er) (mm) shall I talk about |
| 102 | great use of music (erm) I dont know how to say in English \x{d} |
| 103 | who was a a great (erm) . I dont know a great (er) soldier no hes |
| 104 | was a a captain I think I I dont know the degrees (er) of the army |
| 105 | (mm) parts of this movie I dont know the te= the words (er) (em |
| 106 | ake them (mm) less . (mm) I dont know less seriously I think \x{d} |
| 107 | x{d} okay \x{d} \x{d} \x{d} (mm) (eh) I dont know well (mm) w |
| 108 | you can (eh) go and (eh) I dont know read a a book an= and in the |
| 109 | ys and and girls and (mm) I dont know (mm) we have we had a group |
| 110 | d then Italian people and I dont know when we have when we had to |
| 111 | eh) they (eh) . they were I dont know (mm) . (er) they tended to b |
| 112 | our (mm) problems and so I dont know the when I think about them |
| 113 | ear beautiful in order to I dont know (erm) to: to be upset maybe |
| 114 | et maybe from her friends I dont know but shes funny ... (erm) .. |
| 115 | ike the painting and (er) I dont know I \x{d} yes but . the pain |

The second most frequent sequence, *i think that*, will be the subject of cross-corpus comparison in Section 5.1. It should be pointed out, however, that the sequence *I think* is more frequently found in native-speaker speech without the complementiser *that*.

The only other sequence which deserves close attention is *in order to*, as it is quite surprising in a corpus of speech. Figure 5.8 shows a sample of concordance lines for this sequence. In most sentences, however, the subordinator sounds excessively formal and out of place, especially when it is preceded or followed by expressions which are more typical of speech. Here it seems rather illogical that learners choose to use *in order to* instead of *to* and, arguably, in this instance, use of recurrent sequences shows that learners do not adapt their language to the mode of communication. However, if we disregard learners' use of *in order to*, overall, the structural and functional analysis of recurrent 3-word sequences from LINDSEI_IT shows that learner language in speech is characterised by

conversational and casual expressions.

Figure 5.8: Sample Concordances for *in order to* from LINDSEI_IT



| Hit | KWIC |
|---|---|
| 1 | few months (eh) in order . in order to finish my work what I start |
| 2 | ring the summer and (eh) . in order to try to find some Italian cu |
| 3 | at would the best thing to in order to learn the real language the |
| 4 | be more \x{d} yes more hours in order to speak more and more \x |
| 5 | (eh) at the picture maybe in order to control (er) if he is worki |
| 6 | ies of (eh) questions (eh) in order to understand if I was a perso |
| 7 | her people to find friends in order to . (erm) .. get . funny nigh |
| 8 | just wanted to to scare us in order to \x{d} yes and in fact Ive |
| 9 | es shes paying him f= (mm) in order to to have a good a good (er) |
| 10 | nglish not . not German so in order to improve her \x{d} yes \x{ |
| 11 | apher or (er) other people in order to advance in his in his caree |
| 12 | Id like it very much also in order to decide whether to to take m |
| 13 | accept reality \x{d} this is in order to \x{d} yes \x{d} y |
| 14 | so a: (er) I went there to in order to speak English . without my |
| 15 | s a (mm) a journey (er) in in order to visit the U S A so and I wa |
| 16 | nts to to appear beautiful in order to I dont know (erm) to: to be |
| 17 | when I have some time (er) in order to see my boyfriend otherwise |
| 18 | specially when I am abroad in order to see all those (eh) plays wh |
| 19 | o (er) move from our towns in order to study abroad or in other to |
| 20 | she wanted to know hi= him in order to be sure that the baby would |
| 21 | ot live no you have to lie in order to live well in the world \x{ |
| 22 | ertain sense to study them in order to: (eh) (erm) to: to teach th |

## 5.3 Qualitative Cross-corpus Comparison: *think* Clusters and Quantifier Expressions

Since clusters containing the verb *think* and quantifier expressions were found to be among the most frequent sequences in the three learner corpora under analysis, they invite closer attention. Accordingly, they were subjected to additional analysis and the results are presented in the following sections. The results from both the quantitative and the qualitative analyses are taken up and discussed in the light of previous findings on learner English in the conclusive chapter.

### 5.3.1 *Think* Clusters Across the Corpora and Registers

The present section discusses the 3-word clusters found containing the verb *think* in all the three corpora. Clusters were extracted from the three corpora using AntConc©'s cluster function and the verb *think* as a search term. Concordance lines were subsequently analysed in detail in order to shed light on the use of these sequences across the three corpora. Normalised frequencies for 3-word sequences

which have *think* as the node word are presented in Table 5.4; for the purpose of extraction, the cut-off point was set to 20 occurrences for LCC and ICLE_IT and 10 occurrences for LINDSEI_IT, in keeping with previous analyses.

*Think* is normally used at the beginning of an utterance in order to present a personal view on what follows. As can be seen in Table 5.4, two *think* sequences in particular were extracted from the corpora under analysis, namely *i think that*, and *but i think*. Overall, fifteen *think* sequences were extracted from LCC, fourteen from LINDSEI_IT and only four from ICLE_IT. All the ICLE_IT *think* clusters except for the first one have almost negligible frequencies when they are normalised (between 11.4 and 11 per 100,000 words). The normalised frequencies for the other two corpora indicate that the verb *think* is the learners' preferred choice in speech and in chats in a variety of combinations, for example *i think it's* (50.6) and *i think you* (50.6) in LCC, and *and i think* (40.3) and *i don't think* (40.3) in LINDSEI_IT.

Table 5.4: *Think* Clusters in the Three Corpora

| **LCC** | | **ICLE_IT** | | **LINDSEI_IT** | |
|---|---|---|---|---|---|
| **N.F.** | **Cluster** | **N.F.** | **Cluster** | **N.F.** | **Cluster** |
| 130 | i think that | 50.6 | i think that | 166.2 | i think that |
| 50.6 | i think its | 11.4 | but i think | 40.3 | and i think |
| 50.6 | i think you | 11.4 | to think that | 40.3 | i dont think |
| 37 | but i think | 11 | think that the | 38.6 | i think its |
| 31.2 | and i think | | | 28.5 | i think so |
| 23.4 | i think it | | | 26.9 | because i think |
| 21.9 | i think i | | | 26.9 | but i think |
| 20.4 | because i think | | | 21.8 | i think it |
| 18 | think you should | | | 21.8 | i think she |
| 15.6 | i dont think | | | 21.8 | i think they |
| 14.1 | think its a | | | 21.8 | well i think |
| 13.6 | think that you | | | 21.8 | yes i think |
| 11.2 | do you think | | | 16.8 | i think i |
| 11.2 | think that the | | | 16.8 | think that they |
| 10.7 | think that it | | | | |

However, the data show that there are some differences between the two corpora. In LCC *think,* in conjunction with the first person pronoun, is mainly employed to express stance in the present tense. Moreover, LCC includes the

negative and interrogative sequences *i dont think* and *do you think*, whereas no negatives or interrogatives were extracted from ICLE_IT or LINDSEI_IT. Negative stance and interrogative forms reflect the interactivity typical of casual conversation. Negative stance is particularly used to hedge one's opinions (Biber *et al.*, 1999).

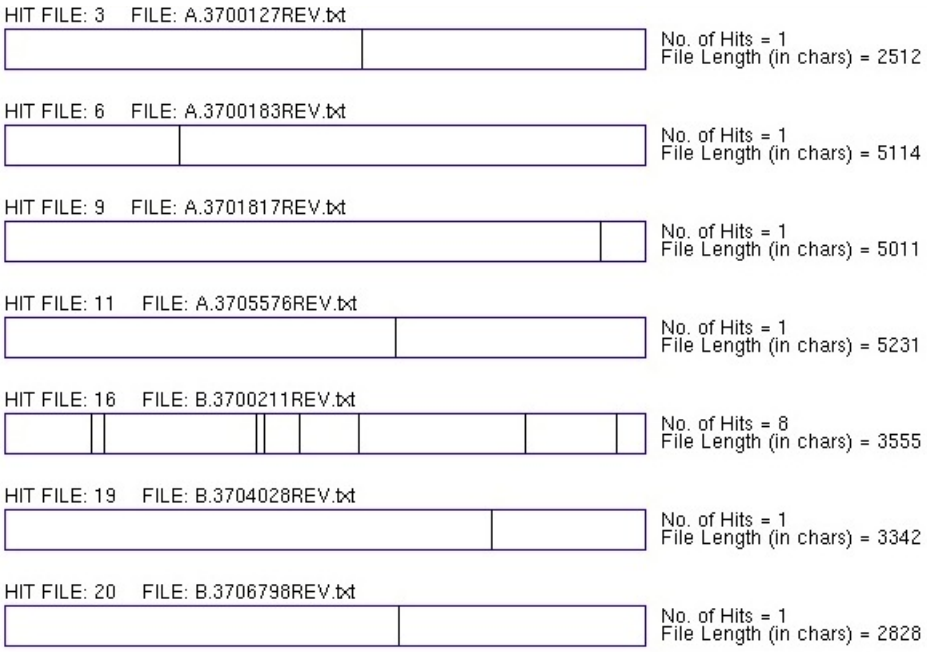
The data from the three corpora reflect differences in terms of the use interactive sequences and of hedging. The lack of interaction and hedging in learner speech has been highlighted by previous research. However, the data from LCC is evidence that the level of interaction in the chat was high. A closer look at *do you think* concordances reveals that the question form is always used in the 4-word expression *what do you think*. In addition, three other sequences from LCC have the pronoun *you* as the object of the verb *think* (namely *i think you*, *think you should*, *think that you*).

The analysis of *think* clusters in LCC reveal a variety of uses of this verb, and the structures in which it is found show a prevalence of the spoken, informal register. The one exception is *i think that*, which is followed by the complementiser, which is more typical of writing than speech. Biber *et al.* (1999) do not count this bundle among the most frequent in English conversation, but it is not reported as typical of academic writing either. As was noted before in the analysis of *in order to*, learners often cling to certain preassembled multi-word expressions.

An in-depth analysis of this sequence shows that in LCC *i think that* is mostly used at the beginning of the sentence and is often preceded by suspension dots, emoticons, and the discourse marker *well*. In some instances it is preceded by the hesitator *mmm* and the response token *oh*. When it is used in the middle of the sentence it is almost always preceded by a connector such as *but*, *and*, or *because*. The corpus data suggest that the sequence is also used for discourse management and as a device to gain thinking time. CMC communication requires a great deal of online processing, which may explain why learners make repeated use of this sequence. It is also interesting to note that the concordance plot for the sequence

shows that while the majority of the learners employ it one to four times in the course of the same chat, fourteen learners repeat it between five and eight times, with two of them including eleven repeats of the sequence. A screenshot of the concordance plot of seven different files from LCC is shown in Figure 5.9. This may well be considered further evidence in support of the view that the sequence is used for discourse management purposes.

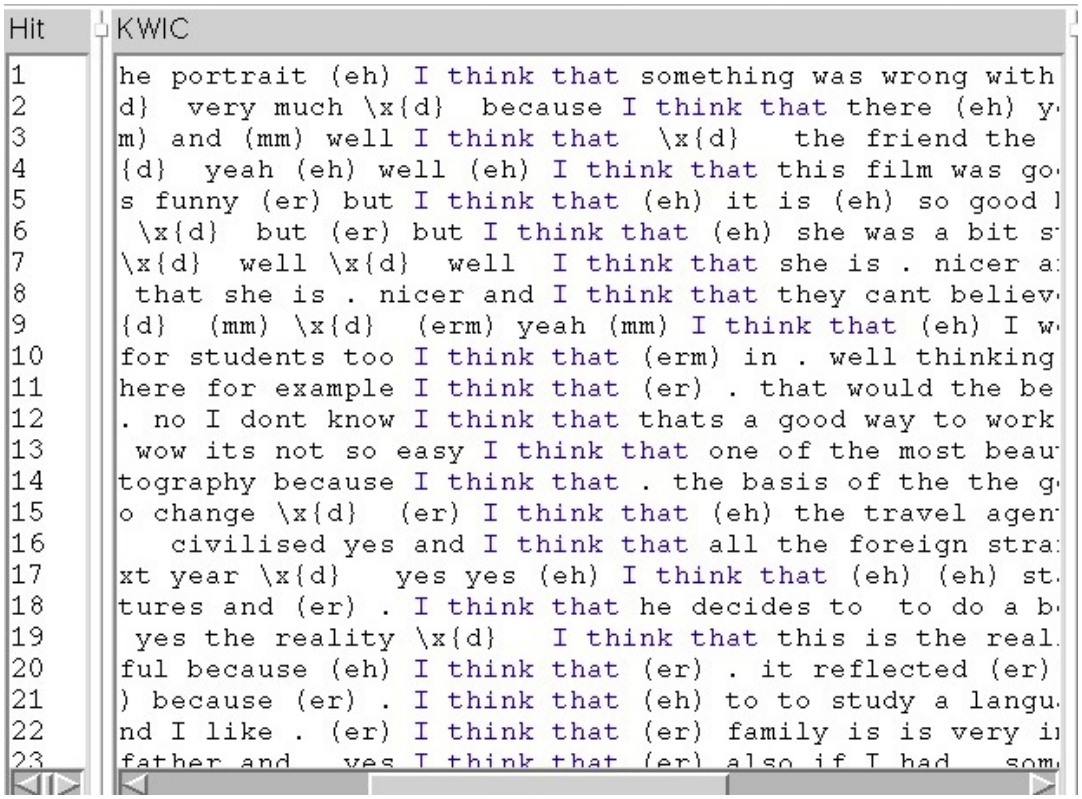
Figure 5.9: LCC: Concordance Plots for *i think that*




The cluster *i think that* is also very frequent in LINDSEI_IT. Figure 5.10 shows sample concordance lines including this expression. As can be noted, the context of use of the sequence is very similar to that brought to light in the LCC data. The sequence is often preceded by hesitation markers, discourse markers, and connectors, which may be interpreted as corroborating evidence that this cluster is in fact used in discourse management, in order to gain processing time, as spoken production requires even more online processing than computer chats.

In ICLE_IT the sequence *i think that* is mostly used at the beginning of the sentence, and in two cases it is in the essay's opening clause. The concordance plot shows that it is used once or twice by most learners, with four learners repeating it four times in the same essay. As was noted before, learners also used

Figure 5.10: Sample Concordance Lines for *i think that* from LINDSEI_IT



| Hit | KWIC |
|---|---|
| 1 | he portrait (eh) I think that something was wrong with |
| 2 | d} very much \x{d} because I think that there (eh) y |
| 3 | m) and (mm) well I think that \x{d} the friend the |
| 4 | {d} yeah (eh) well (eh) I think that this film was go |
| 5 | s funny (er) but I think that (eh) it is (eh) so good |
| 6 | \x{d} but (er) but I think that (eh) she was a bit s |
| 7 | \x{d} well \x{d} well I think that she is . nicer a |
| 8 | that she is . nicer and I think that they cant believ |
| 9 | {d} (mm) \x{d} (erm) yeah (mm) I think that (eh) I w |
| 10 | for students too I think that (erm) in . well thinking |
| 11 | here for example I think that (er) . that would the be |
| 12 | . no I dont know I think that thats a good way to work |
| 13 | wow its not so easy I think that one of the most beau |
| 14 | tography because I think that . the basis of the the g |
| 15 | o change \x{d} (er) I think that (eh) the travel agen |
| 16 | civilised yes and I think that all the foreign stra |
| 17 | xt year \x{d} yes yes (eh) I think that (eh) (eh) st |
| 18 | tures and (er) . I think that he decides to to do a b |
| 19 | yes the reality \x{d} I think that this is the real |
| 20 | ful because (eh) I think that (er) . it reflected (er) |
| 21 | ) because (er) . I think that (eh) to to study a langu |
| 22 | nd I like . (er) I think that (er) family is is very i |
| 23 | father and yes I think that (er) also if I had som |

other sequences to express stance. Although it may be considered unsuitable for the written register, there may be more than meets the eye in learners' use of this sequence. Firstly the fact that most of the ICLE_IT samples are argumentative essays; and secondly, the fact that learners are not experienced writers, and their development as writers of English in academic contexts is not complete.

As for LINDSEI_IT, thirteen out of fourteen *think* clusters include the subject pronoun is *I*, *she* included in one and *they* in two (with normalised frequencies of 21.8, 21.8 and 16.8 respectively). The absence of the pronoun *you*, which is present in LCC, seems to indicate the learners were less involved than in the chat and their output is less interactional. It should be remembered, however, that the transcribed version of an interview lacks all the paralinguistic information such as information on gestures, nodding and eye contact between speakers. These paralinguistic means, together with specific social, cultural and institutional norms should be taken into account when analysing learner speech in terms of interactivity. Paralinguistic signals are notably absent from CMC, which is why learners

need to resort to different means to indicate their participation and interest.

De Cock (2004) noted that the limited use of markers of vagueness in LINDSEI made learner English less involved and more formal; however, it can be seen in LCC that learners can express a greater level of involvement and informality.

A cluster search in LCC reveals that there are fifty-nine different 3-word sequences occurring over 20 times in the corpus which include the pronoun *you* (corresponding to a total of 2,546 tokens). Among them are *if you want*, *I think you, you have to* and *you want to* (all of which have over 100 occurrences). It could be argued that learners were more interactional because the chat itself was devised as an informal exchange and chatting is a mode of communication in which participants are on the same level in terms of power. This cannot be said of teacher-led oral interviews. The same automatic search carried out on LINDSEI_IT extracted only ten sequences in total, nine of which are repetitions of the task rubric and one, namely *you have to*, employs *you* as an impersonal pronoun.

The differences in the use of *you* clusters in the two corpora highlighted above show that the lack of interactivity, which has been interpreted by past research as a failure in terms of register, can in fact be explained by task features and shared context. It should also be borne in mind that the same task in different cultures may be interpreted in different ways, an informal interview with a teacher of English might be devised to be informal in terms of content, but cultural factors will inevitably influence the level of formality and interactivity in terms of language. As they are central in determining learner English features, these issues will be explored in more detail in the next chapter.

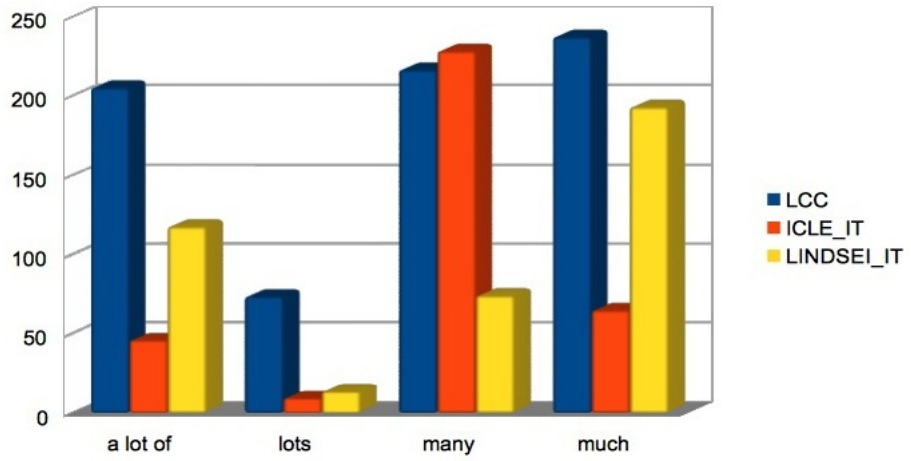
### 5.3.2 Quantifiers Across Corpora and Registers: *a lot of*

Since the quantifier *a lot of* appears at the top of the frequency list in the three corpora, learners' use of quantifiers deserves a closer look. There are a number of expressions that can be used in English to speak about large quantities; con-

sequently, the first analysis was a quantitative comparison of quantifiers across the corpora. The expressions selected for comparison were *a lot of*, *lots*, *many* and *much*, which are reported in Biber *et al.* (1999) as being among the most frequently used in English. Research on corpora of native-speaker speech has shown that the most frequent quantifier in conversation is *much*, followed by *many* and a *lot of*. *Plenty*, *a lot of* and *lots*, which combine with both uncountable and countable plural nouns, are characteristic of casual speech. Academic writing has a preference for *many*, while *a lot of* is used less than 100 times per million words (Biber *et al.*, 1999:275-8).

Figure 5.11 shows normalised frequencies for quantifier expressions for large quantities from LCC, ICLE_IT and LINDSEI_IT. Overall, learners employ *many* more than *a lot of*. *Lots* is the least used and *much* shows higher percentages in LCC and in ICLE_IT.

Figure 5.11: Quantifiers for Large Quantities in the Three Corpora




Consistently with native-speaker use, learner English conversation, exemplified by LINDSEI_IT, shows a prevalence of *much* and *a lot of*. As for *much*, in 84 out of 115 concordance lines it is found premodified by *very*, *so, that* and *too*. In five concordances it is followed by *better*. This is evidence that in learner speech the quantifier *much* is rarely used as a premodifier of a noun. As for *a lot of*, in the 70 concordance lines extracted from LINDSEI_IT, it is used mostly with countable nouns, and only eight uncountable nouns. As can be seen in

the sample concordances in Figure 5.12, at times it is preceded or followed by hesitations and filled pauses. In these cases, it could be argued that *a lot of* is produced by learners in a non-analysed way, as shown by the fact that in some instances quantifier-noun concord is incorrect, as in *we do a lot of written text* and *a lot of scene of violence*.

Figure 5.12: Concordance Lines for *a lot of* from LINDSEI_IT

| | | |
|---|---|---|
| 34 | planes but I dont know because (eh) someone well a lot of people told me that its very difficult to enter a |
| 35 | esting (erm) there is (erm) . (em) the= there are a lot of things there (er) very interesting in Cape Canave |
| 36 | r the actor in fact actors and actress are . gain a lot of money . even if they . didnt work a lot I think |
| 37 | enjoyable because (eh) it (eh) .. (eh) there are a lot of . (erm) yes gags and of course they are focus |
| 38 | ive but when I went to London last year I visited a lot of museum I particularly liked the British Library b |
| 39 | y liked this woman because (eh) she she taught me a lot of things more than just English language so I I had |
| 40 | nd I (erm) (er) and I had the possibility to know a lot of people (eh) from different countries I mean Japan |
| 41 | ic (eh) was very interesting but I I visited (mm) a lot of countries but I dont think that there is a partic |
| 42 | f (mm) marks and (er) during the night I . I made a lot of (eh) of works at home and so I understood that (e |
| 43 | n a hurry about that because I see that there are a lot of possibilities of to work and Im losing Im losing |
| 44 | tion because yes (er) science fiction because a lot of time on the newspapers you can find just the abbr |
| 45 | hat find the si= the signal and (eh) reading with a lot of computers and (eh) (eh) strange (er) (eh) machine |
| 46 | nd this kind this signal (er) with (eh) we we see a lot of digital numbers and (eh) (mm) after some: (er) af |
| 47 | ital numbers and (eh) (mm) after some: (er) after a lot of work they I can understand that this digital numb |
| 48 | uild who builds this machine is United States and a lot of countries . and (er) they makes they make an expe |
| 49 | ll (mm) really not but they had to choose between a lot of candidates and (er) because they wanted to choose |
| 50 | t it was also (eh) nice and (eh) it was there was a lot of irony too and it was a good film because (er) it |
| 51 | pigra lazy Im very lazy and (eh) now I have a lot of things at= t= of things to do but I have no (em) |
| 52 | (eh) (mm) rain wind (eh) we cant we we didnt find a lot of sun= sunny days (erm) I visited everything every |
| 53 | e Champs-Elys√©es Les Invalides (eh) so we have a lot of things to to to s= to to see in a very small (mm) |
| 54 | in Paris (em) is the first fro= (mm) the first of a lot of other towns and places that I would like to visit |
| 55 | year of experience I think Praga because (eh) a lot of people (eh) spoke (mm) tel= (eh) told me that (eh |
| 56 | the city as well is very interesting there are a lot of museums (er) we went to Van Gogh museums which (e |
| 57 | he city is a kind of Venice because (er) there is a lot of water which (er) create a special atmosphere |
| 58 | yes because we no no no but (mm) there were a lot of interesting interesting boys there . specially th |
| 59 | nk its very important to to speak but (erm) we do a lot of written text . summaries (er) and so on but we do |

The quantifier for large quantities that dominates learner English writing, as exemplified by ICLE_IT, is *many*, while *a lot of* and *lots* appear with lower frequencies. It may be argued that, even though these expressions appear in learner writing, overall their usage shows a marked difference between the registers.

The use of *a lot of* in learner essays makes them more informal than standard written English texts, a feature which corroborates existing research on learner writing. Biber *et al.* (1999) report that *a lot of* is rarely found in English academic writing, but it is used in news and in fiction. It should be pointed out that, as stated by Granger *et al.* (2002b), many of ICLE's learners' argumentative essays were written in response to journalistic articles. Therefore, it is hardly surprising that learners' written style employs a quantifier that is found in news articles.

A closer look at the concordance lines of *a lot of* reveals that in ICLE_IT it is mainly used with plural countable nouns, exceptions being *money*, *time*, *coverage*,

*trouble*, *love*, and *imagination*. The most frequently used noun after *a lot of* is *people* (18 out of 104 concordance lines). However, *people* is preceded by the quantifier *many* in 57 concordance lines. Overall, the use of *a lot of* in ICLE_IT seems to be connected to ease of processing: i.e. it is a neutral quantifier whose use does not require grammatical analysis. It should also be noted that the Italian for *people*, *gente*, is uncountable, which may explain why some students choose *a lot of* over *many*. Thus it could be said that in the case of *a lot of* concern for accuracy takes precedence over register appropriateness.

LCC reveals a picture of overuse of all the quantifiers indicating large quantities, which might also be a feature of informal language, which typically resorts to exaggeration. As can be seen in Figure 5.11, with the exception of *lots*, all the quantifiers have very high frequencies. As we saw in the quantitative analysis (4.2.1), the quantifier expression *a lot of* followed by *friends* makes up one of the most frequent 4-word sequences (57 concordances). Additional concordance searches, however, show that concordances of *friend* and *friends* with the quantifier *many* are almost as numerous (53 in total). In some cases the noun *friends* is qualified by some other adjective, as in *Ive a lot of male friends*. Other nouns frequently appearing in conjunction with *a lot of* are *people* (38 concordances) and *things* (35 concordances). The data show that learners do not discriminate in their use of one or the other of these quantifiers. It may be argued, therefore, that the sequence *a lot of* exemplifies how sequences of words tend to become so entrenched in the learners' minds that they are not considered sequences anymore, but rather more like formulae which are retrieved and used as a 'single choice', without grammatical analysis.

A number of conclusions can be drawn from this qualitative look at learners' preferred sequences. Learners were shown to be able to adapt their English to the means of communication, but, at the same time, they were also shown to cling on to preassembled word sequences to carry out some of the functions connected to each register. These research findings will be discussed further in the conclusive

chapter in the light of the claims about learner English made by previous research and of the literature on recurrent sequences in general. As we have seen in sections 2.5-2.5.4, studies of learner English based on corpus data have demonstrated a tendency of learners to repeatedly employ a restricted number of sequences and to show limited awareness of register differences. The data presented in this study provides further evidence of the use of recurrent sequences by learners of English at intermediate to advanced level and only partly corroborates previous research. Chapter Six will discuss research findings and compare the present research study with claims on learner English from previous research. In addition, it will discuss recurrent sequences of words as indicators of learner language processing and the implications of these findings for teaching.

# Chapter Six

# Conclusions

The aim of the present research study was an investigation into the features of the recurrent sequences of words produced by learners of English during an asynchronous chat task. To date, there are very few studies available of learner English CMC, and none regard use of recurrent sequences. Learner corpus research, on the other hand, has already studied learner English writing and speech in terms of lexical bundles and collocations. The present study investigated learner English CMC because of the multiple advantages of using CMC for language production: firstly, it is increasingly part of learners' real-life communicative experiences, and is therefore a study of the use of learner English in a realistic situation; secondly, learners have to deal with a lower processing speed compared to speech; and, thirdly, it is distanced from the speaker by the electronic means and can be monitored while it is produced. These factors explain why CMC brings with it lower levels of anxiety and higher levels of motivation.

The learner English chats were collected in a corpus, and the most frequently occurring sequences of words were extracted by means of a concordancing software and analysed both quantitatively and qualitatively. By means of a detailed description of the most recurrent sequences of words used by learners in the CMC task and of their structural and functional features, the study provides new information which can be added to the current portrait of learner English in the literature.

The following sections (6.1 and 6.2) are a summary and a discussion of the findings of the present research study in the light of current learner corpus research. Section (6.3) expounds the limitations of the present study and its approach, explores future research paths, and discusses the implications of the present findings for teaching practices and teaching materials.

## 6.1 Summary of Findings

Figure 6.1 summarises the findings regarding learners' recurrent sequences extracted automatically from the chat corpus LCC.

Figure 6.1: Features of Learner Recurrent Sequences in CMC

1. Learners use a large stock of recurrent 2-, 3- and 4-word sequence types and tokens, while 5- and 6-word sequences are fewer and have lower frequencies;
2. Learners make considerably frequent use of some 3-word sequences: the top twenty sequences have frequencies between 205.9 and 58.9 per 100,000 words;
3. Recurrent sequences show a prevalence of verb phrases preceded by the first person pronoun;
4. Learner sequences are frequently composed of full structural units;
5. There are very few dependent clause sequences;
6. Learners use recurrent sequences to express personal stance and with referential functions.
7. Learners do not use recurrent sequences to organise discourse;
8. In terms of sequences, learner English in the chats is more spoken than written;
9. Some learners use specific combinations several times in the chat;
10. Most learners show a preference for one or the other sequences which are equivalent, such as *I would like* and *I'd like*.

Sequences of words, especially 3-word ones, are found to be the building blocks of learner English in the chats, analogously to De Cock's (2004) findings regarding learner English speech. However, quantitative evidence from the chats reveals that the types and tokens of 3-word sequences are quite high, even though they are not as many as native speaker ones. Learners' sequences are full structural units and they mostly feature personal pronouns and verb phrases. In terms of functions, learners employ recurrent sequences to express personal stance, while they do not use them to organise discourse. In addition, learners show a preference for specific word combinations which they tend to cling on to.

The main findings of this research both support and complement previous studies of learner English in terms of recurrent sequences. On the one hand, they support findings by De Cock (1998) and De Cock (2004) that learners repeatedly

use the same sequences; on the other hand, however, they show that with a different medium, repetitiveness decreases. The present findings also complement studies about register appropriateness in learner English, since the sequences in the chat corpus were mostly found to be appropriate to the informality and spoken-like quality of CMC. Even if automatic extraction also returned some sequences which may be considered out of place in the spoken register (see *I think that* and *i would like*); close scrutiny of these sequences revealed that some learners use them repeatedly, while others do not, indicating that they are perhaps a personal feature of the speaker's English. Moreover, some repeated sequences seem to have acquired a quasi-formulaic status in the learners' mind and they are produced to ease language processing, as in the case of *a lot of*, which can be used regardless of the following noun.

Overall, learner English in LCC shows a high level of adaptation to the means of communication used. The fact that learners were at ease with the electronic medium is exemplified by the high interactivity of the exchanges and the orientation towards interpersonal interaction and rapport building. These features of learner language are reflected in the large quantity of questions (many of which exhibit non-standard forms), the use of sequences from informal, casual conversation, and the use of emoticons and expressive punctuation. These features are evidence that learners took advantage of the informal and visual nature of CMC.

The next section is an in-depth discussion of findings from the cross-corpus comparison of learner recurrent sequences. It will be shown that comparison provides evidence that learners' use of sequences shows greater adaptation to the register than previously claimed.

## 6.2 Discussion of Findings

This section explores the extent to which learner recurrent sequences from the three corpora analysed show adaptation to the register. The data show that despite vocabulary limitations, learners use word sequences which make their

English suitable to the mode they are using for communication.

Gilquin and Paquot (2008) found that learners' argumentative writing from ICLE shows a clear influence of speech. A comparison of the most recurrent sequences from the Italian subcorpus of ICLE with those of LCC shows that, on the contrary, learners' sequences differ greatly across the two registers. In learner writing from ICLE_IT, recurrent sequences revolve around noun and prepositional phrases, typical of writing; while in LCC personal pronouns and verb phrases dominate recurrent sequences.

Learner writing from ICLE_IT also employs discourse organisers and expressions of stance that are employed in native-speaker writing. A case in point is learners' insistence on the subordinator *in order to*, which is the second most recurrent sequence in ICLE_IT. This expression is found to be appropriate to the register; however, the fact it is repeated so extensively in the essays makes learner writing sound overly repetitive and unnecessarily wordy. The same insistence is found with expressions of stance like *in my opinion*, *point of view*, and the discourse organiser *first of all*.

The repetitive use of these sequences could be explained by the fact that they are novice writers as well as language learners. Excessive writer visibility was shown to be typical of novice writing both in native-speakers (Kibler, 2011) and in learners (Wei, 2009) and the same phenomenon could explain learners' use of sequences such as *I think that* and *in my opinion*.

Sequences from ICLE_IT also show repetitive use of sequences which are not typical of native-speaker writing, such as the quantifier *a lot of* and the stance expression *I think that*. An explanation of this use may come from the fact that learners are not writing academic texts, but argumentative essays on general topics, in some cases they have a newspaper article as starting point. Biber *et al.* (1999) reports that news is less formal than academic writing and the sequence *a lot of*, for example, is found in newspaper articles with much more frequency than in academic prose. Therefore, learners may be trying to imitate journalistic style,

with the result that their prose is less formal than academic writing is generally meant to be.

In summary, evidence from cross-corpus comparison suggests that rather than speech written down, learner writing's use of formal sequences shows they are novice writers trying out the language they are acquiring. In other words, their discourse culture and rhetorical devices do not reflect those of the community of writers in English.

While learners' written production is described as informal and involved (Baker and Chen, 2010), learners' speech was found to be lacking in sequences for interacting and building rapport. Learner speech's recurrent sequences from LINDSEI_IT, however, are characterised by the presence of speech dysfluencies, which essentially indicate encoding problems, and by the highly repetitive use of the stance sequences *i don't know* and *i think that*. Both sequences are among the most frequently repeated in native-speaker speech, the main difference being that the bundle *i think that* is mostly found in native-speaker conversation without the complementiser *that*. With the exception of *in order to*, the other recurrent 3-word sequences in LINDSEI_IT show structural and functional features that are consistent with those of native-speaker speech.

The higher presence of repetitive sequences in LINDSEI_IT may be explained by the influence of the communicative pressure of speech. The phenomenon is also recorded in native-speaker speech (see Biber *et al.*, 1999), where expressions are often repeated to the advantage of processing speed. It should be remembered that LINDSEI is a collection of recorded teacher-led interviews and, in this context, silence and pausing could be evaluated differently than in informal conversation. A case in point is the fact that 3-word sequences from LINDSEI_IT include several repetitions and fillers, which signify the learners interviewed were experiencing encoding problems.

Overall, learner English in LCC is more informal and interactional than learner speech from LINDSEI_IT. The explanation for this finding may well be that

learners' output becomes interactional when the task requires interaction. In the chat, learners were participating in a speech event as peers, while in the interviews they were carrying out tasks which required much less interaction.

In LCC, learner sequences show additional features that are worth discussing. The level of formality exhibited in sequences in LCC is decidedly low. The quantity and the type of output produced in the chats clearly shows that the task did not produce anxiety in learners; on the contrary, generally speaking, they were greatly involved, as evidenced by the length of chats and the high level of interactivity. Data analysis shows that accuracy was not the learners' main concern in the chats, in fact, they used non-standard capitalisation (*you MUST go*), abbreviations *(if u r)*, non-standard grammar (e.g. *sounds nice!*), and expressions from casual speech and informal writing (e.g. *wanna*, *lemme*). In CMC all these features are used for rapport building (see Herring, 2012), the same functions carried out by question forms and emoticons, which are both employed extensively throughout the corpus. In this context, well-rehearsed routines such as *i really like* or *nice to meet you*, which can be produced effortlessly, have the immediate effect of showing willingness to participate, to build rapport and mark involvement.

Summing up, the cross-corpus comparison showed that recurrent sequences in the three corpora show distinctive features: sequences from LCC have a prevalence of *I* subject plus verb phrase, sequences from ICLE_IT revolve around noun and prepositional phrases, typical of writing, while the most distinctive feature of sequences from LINDSEI_IT is the presence of dysfluencies typical of speech and the repetitive use of informal sequences. Conversely, comparisons across communication modes has revealed that two sequences are used frequently by learners in writing, speech and CMC, namely sequences including the verb *think*, and the quantifier sequence *a lot of*. On the one hand the use of these expressions seem to substantiate the claims made by previous research that learners employ the same recurrent sequences regardless of the register; on the other hand, learners'

failures in terms of the register of recurrent sequences may be explained by the ease of processing and the principle of minimum effort: learners may be relying on recurrent sequences because they can produce them holistically, freeing their working memory and gaining thinking time in the process.

After the analysis and discussion of findings from the present research, we may safely conclude that the fact that learner's use of recurrent sequences changes with the means of communication is ample demonstration that learners do register frequency differences in the input, be it spoken, written or CMC, and produce language accordingly. While in the past learner input was mainly teaching materials, teacher talk and reading texts, learner English from LCC shows that learners nowadays have many other occasions to encounter and practice the language and this is perhaps contributing to greater register awareness. However, since teaching practice and materials clearly have a great influence on learner output, one of the tasks that teachers should be aware of, especially at upper-intermediate and advanced level, should be to make register differences clear and train students to identify the relationship between recurrent sequences of words and the different means of communication.

## 6.3 Limitations of the Study and Future Research

While the strength of the present research lies in its bottom-up approach to the study of recurrent sequences in learner English across registers, clearly the recurrent sequences identified by means of automatic extraction do not provide a complete picture of learner English CMC. As demonstrated by the analysis of concordance lines from the three corpora in Chapters Four and Five, an abundance of facts about learner English is awaiting discovery by means of more qualitative analysis of corpus data.

Further studies could delve into the contexts in which the recurrent sequences

are employed and find information about grammar, collocates, syntax, and so on. These analyses of LCC texts could even promote the creation of exercises for learners at upper-intermediate and advanced level, which would better prepare them to use English across registers in their future professions. Qualitative research could also complement the findings of this study and reveal aspects of learner English which have escaped the present analysis. In particular, for example, regarding the linguistic means used by learners for hedging, or to express vagueness and contribute to paint an ever more detailed picture of learner English.

It is hoped that the 200,000-word corpus collected for the present research is a stimulus for future investigations into the features of learner English produced by means of CMC. Future research projects could make use of LCC to correlate learner English features and the learner data collected in the survey, which did not find space in the present thesis. The information collected in the survey regards the language background of the learners and their language learning experiences abroad and could be employed to create groups of learner chats to be compared using different variables. In addition, future research studies could seek correlations between the familiarity with CMC shown by learners' use of non-standard capitalisation, repetition of letters and emoticons and other measures of learner language pragmatic development. Further qualitative research into LCC could also delve into specific chats and compare them in detail. For example, since in the course of the present research it was found that some learners employ the same sequences many times in the course of the same chat, while others do not, it would clearly be worth exploring possible correlations between the repeated use of the same sequences and measures of fluency and proficiency. These studies would provide scientific evidence of the results of local teaching practices, which would be useful for the creation of ad hoc teaching materials.

In sum, it is hoped that the present analysis of the LCC represents the first step of a line of research studies regarding this new communication modality in order to provide ever more evidence of learner language processing and use which

could then usefully inform teaching practice and learning materials.

# References


Abrams, Z.I. (2003). 'The Effect of Synchronous and Asynchronous CMC on Oral Performance in German'. *The Modern Language Journal*, 87 (2), 157–167.

Ädel, A. and Erman, B. (2012). 'Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: A Lexical Bundles Approach'. *English for Specific Purposes*, 31 (2), 81–92.

Ädel, A. (2011). 'Rapport building in student group work'. *Journal of Pragmatics*, 43 (12), 2932–2947.

Aijmer, K. (1996). *Conversational Routines in English : Convention and Creativity*. Studies in language and linguistics, London: Longman.

—— (2009a). *Corpora and Language Teaching*. Studies in Corpus Linguistics, Amsterdam/Philadelphia: John Benjamins.

—— (2009b). '"So er I just sort I dunno I think it's just because...": A Corpus Study of I don't know and dunno in Learners' Spoken English'. *Language and Computers*, 68:1, 151–168.

—— (2011). '"Well I'm not sure I think": The Use of well by Non-native Speakers'. *International Journal of Corpus Linguistics*, 16 (2), 231–254.

Aijmer, K. (2004). 'Pragmatic Markers in Spoken Interlanguage'. *Nordic Journal of English Studies*, 3:1, 173–190.

Allen, D. (2009). 'Lexical Bundles in Learner Writing: An Analysis of Formulaic Language in the ALESS Learner Corpus'. *Komaba Journal of English Education*, 1, 105–127.

Altenberg, B. (1998). 'On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations'. In A.P. Cowie (ed.) *Phraseology: Theory, Analysis, and Applications*, pp. 101–122, Oxford: Oxford University Press.

Baker, P. and Chen, Y. (2010). 'Lexical Bundles in L1 and L2 Academic Writing.' *Language Learning and Technology*, 14 (2), 30–49.

Barfield, A. and Gyllstad, H. (2009). *Researching Collocations in Another Language: Multiple Interpretations*. Basingstoke/New York: Palgrave Macmillan.

Bednarek, M. and Bublitz, W. (2007). 'Enjoy!: The (Phraseological) Culture of Having Fun'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 109–136, Berlin: Walter de Gruyter.

—— (2006). 'University Language: A Corpus-Based Study of Spoken and Written Registers'.
**URL:** *http://books.google.it/books?id=-2zqpWi19h4C*

—— (2009). 'A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing'. *International Journal of Corpus Linguistics*, 14 (3), 275–311.

Biber, D. and Barbieri, F. (2007). 'Lexical bundles in university spoken and written registers'. *English for Specific Purposes*, 26:3, 263–286.

Biber, D., Johansson, S. *et al.* (1999). *Longman grammar of spoken and written English*. Harlow :: Longman,.

Biber, D., Conrad, S. *et al.* (2004). 'If you look at... : Lexical Bundles in University Teaching and Textbooks'. *Applied Linguistics*, 25:3, 371–405.

Block, D. and Cameron, D. (2002). *Globalization and Language Teaching*. Taylor & Francis Group.

Brand, C. and Götz, S. (2011). 'Fluency versus accuracy in advanced spoken learner language: A multi-method approach'. *International Journal of Corpus Linguistics*, 16:2, 255–275.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Bygate, M., Skehan, P. *et al.* (2001). *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Applied Linguistics and Language Study, Longman.

Callies, M. (2008). 'Easy to Understand but Difficult to Use? Raising Constructions and Information Packaging in the Advanced Learner Variety'. In L. up contrastive, learner corpus research. G. Gilquin S. Papp, and D. M.B. (eds.) *Linking up Contrastive and Learner Corpus Research*, pp. 201–226, Amsterdam/New York: Rodopi.

—— (2009). *Information Highlighting in Advanced Learner English: The Syntax-Pragmatics Interface in Second Language Acquisition*. John Benjamins Publishing Company.

Callies, M. and Zaytseva, E. (2011). 'The Corpus of Academic Learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties'. In T. Hedeland H. Schmidt and K.e. Worner (eds.) *Multilingual Resources and Multilingual Applications (Hamburg Working Papers in Multilinguialism*, vol. B96, pp. 51–56.

Campoy, M.C. and Luzon, M.J. (2007). *Spoken Corpora in Applied Linguistics*. Linguistic Insights, v. 51 1424-8689, Peter Lang.

Candlin, C. and Hyland, K. (1999). *Writing: Texts, Processes, and Practices*. Harlow: Longman.

Carlsen, C. (2012). 'Proficiency Level: a Fuzzy Variable in Computer Learner Corpora'. *Applied Linguistics*, 33:2, 161–183.

Cobb, T. (2003). 'Analyzing late interlanguage with learner corpora: Quebec replications of three European studies.' *The Canadian Modern Language Review*, 59:3, 394–423.

—— (1996). 'Electronic Language: a New Variety of English'. In S.C. Herring (ed.) *Computer Mediated Communication. Linguistic, Social and Cross Cultural Perspectives*, pp. 13–28, Amsterdam/Philadelphia: John Benjamins.

Columbus, G. (2010). *Perspectives on Formulaic Language: Acquisition and Communication*, chap. Processing MWUs: Are MWU Subtypes Psycholinguistically Real?, pp. 194–212. London/New York: Continuum.

Conklin, K. and Schmitt, N. (2008). 'Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?' *Applied Linguistics*, 29:1, 72–89.

Connor, U. and Upton, T. (2004). *Applied Corpus Linguistics: A Multidimensional Perspective*. New York: Rodopi.

Cortes, V. (2004). 'Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology'. *English for Specific Purposes*, 23:4 (4), 397 – 423.

Cowie, A.P. (1998). *Phraseology: Theory, Analysis, and Application*. Oxford linguistics, Oxford: Oxford University Press.

Crystal, D. (2003). *English as a Global Language*. Cambridge: Cambridge University Press.

—— (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Danet, B. and Herring, S. (2007). *The Multilingual Internet: Language, Culture, and Communication Online*. New York: Oxford University Press.

De Cock, S., Granger, S. *et al.* (1998). *An Automated Approach to the Phrasicon of EFL Learners*, pp. 67–79. London/New York: Addison Wesley Longman.

De Cock, S. (1998). 'A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-native Speakers of English'. *International Journal of Corpus Linguistics*, 3 (1), 59–80.

—— (2004). 'Preferred Sequences of Words in NS and NNS Speech'. *Belgian Journal of English Language and Literature (BELL) New Series*, 2, 225–246.

Deutschmann, M., Ädel, A. *et al.* (2009). 'Introducing Mini-McCALL: A Pilot Version of the Mid-Sweden Corpus of Computer-Assisted Language Learning'. *ICAME Journal*, 33, 21–44.

Dörnyei, Z. and Ushioda, E. (2009). *Motivation, Language Identity and the L2 Self*. Second Language Acquisition, Bristol: Multilingual Matters.

Dresner, E. and Herring, S.C. (2010). 'Functions of the Nonverbal in CMC: Emoticons and Illocutionary Force'. *Communication Theory*, 20 (3), 249–268.

Dresner, E. (2005). 'The Topology of Auditory and Visual Perception, Linguistic Communication, and Interactive Written Discourse'. *language@internet*, 2 (2). **URL:** *www.languageatinternet.org/articles*

Durrant, P. and Doherty, A. (2010). 'Are High-frequency Collocations Psychologically Real? Investigating the Thesis of Collocational Priming'. *Corpus Linguistics and Linguistic Theory*, 6 (2), 125–155.

Ellis, N.C. and Larsen-Freeman, D. (2009). *Language as a Complex Adaptive System*. Language learning, v. 59, Chichester: Wiley-Blackwell.

Ellis, N.C. (2002). 'Frequency Effects in Language Processing'. *Studies in Second Language Acquisition*, 24 (2), 143–188.

Ellis, N.C., Simpson-Vlach, R. *et al.* (2008). 'Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL'. *TESOL Quarterly*, 42 (3), 375–396.

Erman, B. (2007). 'Cognitive Processes as Evidence of the Idiom Principle'. *International Journal of Corpus Linguistics*, 12 (1), 25–53.

Fernando, C. (1996). *Idioms and Idiomaticity*. Describing English language, Oxford: Oxford University Press,.

Foss, P. (2009). 'Constructing a Blog Corpus for Japanese Learners of English'. *The Jalt CALL Journal*, 5 (1), 65–76.

Francis, G. (1993). 'A Corpus-Driven Approach to Grammar: Principles, Methods and Examples'. In M. Baker, F. Gill, and E.T. Bonelli (eds.) *Text and Technology: In honour of John Sinclair*, pp. 137–156, John Benjamins.

Gerbig, A. and Shek, A. (2007). 'The Phraseology of Tourism: A Central Lexical Field and its Cultural Construction'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 303–322, Berlin: Walter de Gruyter.

Ghadessy, M., Henry, A. *et al.* (2001). *Small Corpus Studies and Elt: Theory and Practice*. Amsterdam/Philadelphia: J. Benjamins Publishing Company.

Gilquin, G. (2000). 'The Integrated Contrastive Model. Spicing up your data'. *Languages in Contrast*, 3:1, 95–123.

—— (2005). 'To Take or not to Take Phraseology into Account. The Place of Multi-word Expressions in Corpus Data and Experimental Data'. In F.M. C. Cosme C. Gouverneur and M. Paquot (eds.) *Proceedings of the Phraseology 2005 Conference, Louvain-la-Neuve, 13-15 October 2005*, pp. 165–168.

—— (2009). *Corpora and Experimental Methods*. Corpus linguistics and linguistic theory, Mouton De Gruyter.

Gilquin, G. and Cock, S.D. (eds.) (2011). *Errors and Dysfluencies in Spoken Corpora*, John Benjamins.

Gilquin, G. and Granger, S. (2011). 'From EFL to ESL: Evidence from the International Corpus of Learner English'. In J. Mukherjee and M. Hundt (eds.) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, pp. 55–78, Amsterdam/Philadelphia: John Benjamins.

Gilquin, G. and Paquot, M. (2007). 'Spoken features in learner academic writing: identification, explanation and solution'. In S.H..P.D.e. M. Davies P. Rayson (ed.) *Proceedings of the Fourth Corpus Linguistics Conference CL2007, University of Birmingham, 27-30 July 2007*, University of Birmingham, UK.

—— (2008). 'Too chatty: Learner academic writing and register variation'. *English Text Construction*, 1 (1), 41–61.

Gilquin, G., Papp, S. *et al.* (eds.) (2008). *Linking up contrastive and learner corpus research*, Language and computers, no. 66 0921-5034, and International Contrastive Linguistics Conference. (4th : 2005 : Santiago de Compostela, Spain), Amsterdam [etc.] : 2008: Rodopi.

Gilquin, G., De Cock, S. *et al.* (2010). *Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Presses universitaires de Louvain.

Gilquin, G., Granger, S. *et al.* (2007). 'Learner corpora: The missing link in EAP pedagogy'. *Journal of English for Academic Purposes*, 6 (4), 319–335.

Gong, W. (2005). 'English in computer-mediated environments: a neglected dimension in large English corpus compilation'. In *Proceedings from the Corpus Linguistics Conference Series*.

Götz, S. and Schilk, M. (2011). 'Formulaic Sequences in Spoken ENL, ESL and EFL'. In J. Mukherjee and M. Hundt (eds.) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, pp. 79–100, Amsterdam/Philadelphia: John Benjamins.

Goutsos, D. (2005). 'The interaction of generic structure and interpersonal relations in two-party e-chat discourse'. *language@internet*, 2:3.

Granger, S. (1996). 'From CA to CIA and Back: an integrated approach to computerized bilingual and learner corpora'. In B.A. K. Aijmer and M. Johansson (eds.) *Languages in Contrast. Papers from a Symposium on Text-based Crosslinguistic Studies. Lund 4-5 March 1994.*, pp. 37–51, Lund: Lund University Press.

Granger, S. and de Louvain, U.C. (2009). *The international corpus of learner English*. Louvain-la-Neuve, Belgium: Universite catholique de Louvain, Centre for English Corpus Linguistics„ version 2. ed.

Granger, S. and Paquot, M. (2008). 'Disentangling the Phraseological Web'. In S. Granger and F. Meunier (eds.) *Phraseology: An interdisciplinary Perspective*, pp. 27–49, Amsterdam & Philadelphia: John Benjamins.

Granger, S., Hung, J. *et al.* (2002a). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. John Benjamins Pub.

Granger, S., Dagneaux, E. *et al.* (2002b). *International Corpus of Learner English : Version 1.1 ; Handbook and CD-ROM*. Louvain-la-Neuve: Pr. Univ. de Louvain.

Granger, S. (1998). *Learner English on computer*. Studies in language and linguistics, London/New York: Addison Wesley Longman.

—— (2003). 'The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research'. *TESOL Quarterly*, 37 (3), 538–546.

—— (2004a). *Applied Corpus Linguistics: A Multidimensional Perspective*, chap. Computer Learner Corpus Research: Current Status and Future Prospects, pp. 123–145. Amsterdam & Atlanta: Rodopi.

—— (2004b). 'Computer learner corpus research: current status and future prospects'. In *Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam & Atlanta: Rodopi*, pp. 123–145.

Granger, S. and Meunier, F. (2008). *Phraseology : an interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins.

Granger, S. and Rayson, P. (1998). *Learner English on Computer*, chap. Automatic Lexical Profiling of Learner Texts, pp. 119–131. London & New York: Addison Wesley Longman.

Gries, S.T. (2008). *Phraseology and linguistic theory: a brief survey.*, vol. Phraseology: an interdisciplinary perspective, pp. 3–25. Amsterdam/Philadelphia: John Benjamins.

Herring, S.C. (2010). 'Computer-mediated conversation: Introduction and overview'. *Language@Internet*, 7, article 2.

—— (2012). 'Grammar and electronic communication'. In C. Chapelle (ed.) *Encyclopedia of Applied Linguistics*, Hoboken, NJ:: Wiley-Blackwell.

—— (1996). *Computer-mediated Communication : Linguistic, Social and Cross-cultural Perspectives*. Pragmatics & beyond. New series ; 39, Amsterdam: John Benjamins.

—— (2002). 'Computer-mediated communication on the internet'. *Annual Review of Information Science and Technology*, 36 (1), 109–168.

Howarth, P. (1998). 'Phraseology and Second Language Proficiency'. *Applied Linguistics*, 19 (1), 24–44.

Hulstijn, J.H. (2007). 'The Shaky Ground beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency'. *The Modern Language Journal*, 91 (4), pp. 663–667.

Hyland, K. and Milton, J. (1997). 'Qualification and certainty in L1 and L2 students writing'. *Journal of Second Language Writing*, 6, 183–205.

Hyland, K. (2008a). 'Academic clusters: text patterning in published and post-graduate writing'. *International Journal of Applied Linguistics*, 18 (1), 41–62.

—— (2008b). 'As can be seen: Lexical bundles and disciplinary variation'. *English for Specific Purposes*, 27 (1), 4–21.

Jucker, A., Schreier, D. *et al.* (2009). *Corpora: Pragmatics and Discourse : Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. Language and Computers: Studies in Practical Linguistics, Amsterdam/New York: Rodopi.

Kachru, B.B. (1985). 'Standards, codification and sociolinguistic realism: the English language in the outer circle.' In R. Quirk and H.G. Widdowson (eds.) *English in the World: Teaching and Learning the Language and Literatures*, pp. 11–30, Cambridge: Cambridge University Press.

Kibler, A. (2011). '"I write it in a way that people can read it": How Teachers and Adolescent L2 Writers Describe Content Area Writing'. *Journal of Second Language Writing*, 20 (3), 211–226.

Kirkpatrick, A. (2010). *The Routledge Handbook of World Englishes*. Routledge Handbooks in Applied Linguistics, Taylor & Francis.

Kjellmer, G. (1994). *A dictionary of English Collocations: based on the Brown Corpus*, vol. 1. Clarendon Press.

Kötter, M. (2003). 'Negotiation of Meaning and Codeswitching in Online Tandems'. *Language Learning and Technology*, 7 (2), 145–172.

Kuiper, K. (2004). 'Formulaic Performance in Conventionalised Varieties of Speech'. In N. Schmitt (ed.) *Formulaic Sequences: Acquisition, Processing, and Use*, pp. 37–54, Amsterdam/Philadelphia: John Benjamins.

Kuiper, K. and Lin, D.T.G. (1989). 'Cultural Congruence and Conflict in the Acquisition of Formulae in a Second Language'. In O. Garcia and R. Otheguy (eds.) *English Across Cultures: Cultures Across English*, pp. 281–304, Berlin: Walter de Gruyter.

Laane, M.A. (2011). 'Lexical Bundles in Engineering Research Articles'. *in Proceedings, 10th International Symposium. Topical problems in the Field of Electrical and Power Engineering.*, pp. 72–75.

Larsen-Freeman, D. and Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.

Lee, P. (2007). 'Formulaic language in cultural perspective'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 471–496, Berlin: Walter de Gruyter.

Li, J. and Schmitt, N. (2009). 'The acquisition of lexical phrases in academic writing: A longitudinal case study'. *Journal of Second Language Writing*, 18:2, 85–102.

Lin, P.M.S. (2010). 'The Phonology of Formulaic Sequences: A Review'. In D. Wood (ed.) *Perspectives on Formulaic Language: Acquisition and Communication*, pp. 174–193, London/New York: Continuum.

—— (2012). 'Sound Evidence: The Missing Piece of the Jigsaw in Formulaic Language Research'. *Applied Linguistics*, 33 (3), 342–347.

Lüdeling, A. and Kytö, M. (2008). *Corpus Linguistics : an International Handbook*, vol. 29. Walter de Gruyter.

Macfadyen, L.P. (2006a). 'The Prospects for Identity and Community in Cyberspace'. In C. Ghaoui (ed.) *Encyclopedia of Human Computer Interaction*, pp. 471–478, Hershey, PA: Information Science Reference.

Macfadyen, L.P. (2006b). 'Internet-Mediated Communication at the Cultural Interface'. In *Encyclopedia of Human Computer Interaction*, pp. 373–380, IGI Global.

Mair, C. (2007). 'Varieties of English Around the World: Collocational and Cultural Profiles'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 437–470, Berlin: Walter de Gruyter.

Mair, C., Hundt, M. *et al.* (2000). *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*. Language and computers, Rodolpi.

McCrostie, J. (2008). 'Writer visibility in EFL learner academic writing: A corpus-based study'. *ICAME Journal*, 32, 97–114.

McEnery, T. and Hardie, A. (2012). *Corpus linguistics : method, theory and practice*. Cambridge textbooks in linguistics, Cambridge/New York: Cambridge University Press.

McEnery, T., Xiao, R. *et al.* (2006). *Corpus-based Language Studies : an Advanced Resource Book*. Routledge applied linguistics, London/New York: Routledge.

Meunier, F. and Granger, S. (2008). *Phraseology in foreign language learning and teaching*. John Benjamins.

Minugh, D.C. (2008). 'The College Idiom: Idioms in the COLL Corpus'. *ICAME Journal*, 32, 115–138.

Moon, R. (1998). *Fixed expressions and idioms in English : a corpus-based approach*. Oxford studies in lexicography and lexicology, Oxford: Clarendon Press.

Mukherjee, J. and Hundt, M. (2011). *Exploring Second-language Varieties of English and Learner Englishes : Bridging a Paradigm Gap*, vol. 44 of *Studies in corpus linguistics*. Amsterdam/Philadelphia: John Benjamins.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse.* Amsterdam/Philadelphia :: John Benjamins Publishing Company,.

Murray, D.E. (2000). 'Protean Communication: The Language of Computer-Mediated Communication'. *TESOL Quarterly*, 34 (3), 397–421.

Nattinger, J.R. and DeCarrico, J.S. (1992). *Lexical Phrases and Language Teaching*. Oxford applied linguistics, Oxford/New York: Oxford University Press.

Neff, J., Ballesteros, F. *et al.* (2004). 'Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora'. *Language and Computers*, 52 (1), 73–89.

Nekrasova, T.M. (2009). 'English L1 and L2 Speakers' Knowledge of Lexical Bundles'. *Language Learning*, 59, 647–686.

Nesi, H. and Basturkmen, H. (2006). 'Lexical Bundles and Discourse Signalling in Academic Lectures'. *International Journal of Corpus Linguistics*, 11 (3), 283–304.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Advances in Consciousness Research, J. Benjamins Publishing Company.

—— (2009). 'Co-selection phenomena across New Englishes Parallels (and differences) to foreign learner varieties'. *English World-Wide*, 30 (1), 1–26.

Paolillo, J. (1999). 'The Virtual Speech Community: Social Network and Language Variation on IRC'. *Journal of Computer-Mediated Communication*, 4 (4).

Pawley, A. (2007). 'Developments in the Study of Formulaic Language since 1970: A Personal View'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 3–45, Berlin: Walter de Gruyter.

Pawley, A. and Syder, F. (1983). *Two Puzzles for Linguistic Theory: Native-like Selection and Native-like Fluency*.

Peeters, B. (2007). 'Australian perceptions of the weekend: Evidence from collocations and elsewhere'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 79–108, Berlin: Walter de Gruyter.

Pravec, N.A. (2002). 'Survey of Learner Corpora'. *ICAME Journal*, 26, 81–114.

Rica Peromingo, J. (2010). 'The use of lexical bundles in the written production of Spanish EFL university students'. *Applied Linguistics for Specialised Discourse*, pp. 1–7.

Römer, U. (2005). *Progressives, Patterns. Pedagogy: A Corpus-driven Approach to English Progressive Forms, Functions, Contexts, and Didactics*. J. Benjamins Publishing Company.

—— (2007). 'Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals'. In *Volk-Birke, Sabine & Julia Lippert (eds.). Anglistentag 2006 Halle. Proceedings. Trier: Wissenschaftlicher Verlag Trier.*, pp. 355–363.

—— (2009). 'English in academia: Does nativeness matter?' *Anglistik: International Journal of English Studies*, 20 (2), 89–100.

Sauro, S. and Smith, B. (2010). 'Investigating L2 Performance in Text Chat'. *Applied Linguistics*, 31 (4), 554–577.

Schauer, G. (2009). *Interlanguage Pragmatic Development: The Study Abroad Context*. London/New York: Continuum.

Schiffrin, D., Tannen, D. *et al.* (2003). *The handbook of discourse analysis*. Blackwell handbooks in linguistics, Malden/Oxford: Blackwell.

Schmied, J. (2011a). 'Academic Writing and New Englishes: Unifying the Contrasts'. *Discourse and Interaction*, 4.

—— (2011b). 'Academic Writing and New Englishes: Unifying the Contrasts'. *Brno Studies in English*, 37, 39–47.

Schmied, J. and Haase, C. (2011). *Academic Writing in Europe: Empirical Perspectives*. Cuvillier Verlag.

Schmitt, N. and McCarthy, M. (1998). *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

—— (2004). *Formulaic sequences : Acquisition, Processing, and Use*. Language learning and language teaching, v. 9 1569-9471, Amsterdam/Philadelphia: John Benjamins.

Schönefeld, D. (2007). 'Hot, Heiss, and Gorjachij: A Case Study of Collocations in English, German, and Russian'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 137–177, Berlin: Walter de Gruyter.

Sharifian, F. and Palmer, G. (2007). *Applied Cultural Linguistics: Implications for Second Language Learning and Intercultural Communication*. John Benjamins.

Simpson, R. and Mendis, D. (2003). 'A Corpus-Based Study of Idioms in Academic Speech'. *TESOL Quarterly*, 37 (3), 419–441.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Describing English language, Oxford University Press.

—— (1999). 'A Way with Common Words'. In H. Hasselgard and S. Oksefjell (eds.) *Out of Corpora: Studies in Honour of Stig Johansson*, vol. 26, pp. 157–180, Rodopi.

Sinclair, J. and Carter, R. (2004). *Trust The Text: Language, Corpus and Discourse*. New York: Routledge.

Sinclair, J. (1996). 'The Empty Lexicon'. *International Journal of Corpus Linguistics*, 1 (1), 99–119.

Siyanova-Chanturia, A., Conklin, K. *et al.* (2011). 'Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers'. *Second Language Research*.

Skandera, P. (2007). *Phraseology and Culture in English*. Topics in English Linguistics [TiEL] Series, Amsterdam: Walter De Gruyter.

Stubbs, M. (2002). 'Two quantitative methods of studying phraseology in English'. *International Journal of Corpus Linguistics*, (7).

—— (2007). 'An Example of Frequent English Phraseology: Distributions, Structures and Functions'. *Language and Computers*, 62 (1), 89–105.

—— (in press). 'Sequence and order: the neo-Firthian tradition of corpus semantics.' In J.E..S.O. In H. Hasselgard (ed.) *Corpus Perspectives on Patterns of Lexis*, Amsterdam: Benjamins.

—— (2009). 'Memorial Article: John Sinclair (1933-2007)'. *Applied Linguistics*, 30 (1), 115–137.

Thorne, S.L. and Black, R.W. (2007). 'Language and Literacy Development in Computer-Mediated Contexts and Communities'. *Annual Review of Applied Linguistics*, 27, 133–160.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Studies in Corpus Linguistics, J. Benjamins.

Tono, Y. (2003). 'Learner corpora: Design, development and applications'. In *Lancaster University: University Centre for Computer Corpus Research on Language*, pp. 800–809.

Tremblay, A., Derwing, B. *et al.* (2011). 'Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks'. *Language Learning*, 61 (2), 569–613.

Tucker, G. and Corson, P. (2008). *Encyclopedia of Language and Education*, vol. 4. Springer.

Wei, N. (2009). 'On the Phraseology of Chinese Learner Spoken English: Evidence of Lexical Chunks from COLSEC'. *Language and Computers*, 68 (1), 271–296.

—— (2007a). 'Reasonably well: Natural Semantic Metalanguage as a Tool for the Study of Phraseology and its Cultural Underpinnings'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 49–78, Berlin: Walter de Gruyter.

—— (2007b). 'Reasonably well: Natural Semantic Metalanguage as a tool for the study of phraseology and of its cultural underpinnings'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 49–78, Berlin.

—— (2009). 'Exploring English Phraseology with Two Tools'. *Journal of English Linguistics*, 37 (2), 101–129.

Wiktorsson, M. (2001). 'Register Differences between Prefabs in Native and EFL English'. *The Department of English in Lund: Working Papers in Linguistics*, 1, 85–94.

Wolf, H.G. and Polzenhagen, F. (2007). 'Fixed Expressions as Manifestations of Cultural Conceptualizations: Examples from African Varieties of English'. In P. Skandera (ed.) *Phraseology and Culture in English*, pp. 399–436, Berlin: Walter de Gruyter.

—— (2010b). *Perspectives on Formulaic Language: Acquisition and Communication*. Bloomsbury.

Wray, A. (2005). *Formulaic Language and the Lexicon*. Cambridge University Press.

—— (2000). 'Formulaic sequences in second language teaching: principle and practice'. *Applied Linguistics*, 21 (4), 463–489.

—— (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

—— (2008). *Formulaic Language : Pushing the Boundaries*. Oxford applied linguistics, Oxford: Oxford University Press.

Wray, A. and Perkins, M.R. (2000). 'The Functions of Formulaic Language: an Integrated Model'. *Language & Communication*, 20 (1), 1–28.

Wulff, S. (2008). *Rethinking idiomaticity: a usage-based approach*. Research in corpus and discourse, Continuum.

Wynne, M. (2005). *Developing Linguistic Corpora : a Guide to Good practice*. Oxford: Oxbow Books on behalf of the Arts and Humanities Data Service.

Yus, F. (2011). *Cyberpragmatics: Internet-Mediated Communication in Context*. Pragmatics & beyond new series, Amsterdam/Philadelphia: John Benjamins.

# Appendix I

## Cherry's Chat Turns:

1 Hi! My name's Cherry and I'm from the US. I'm here on Blackboard to chat with you. Thanks for coming, I was waiting for someone to chat with... Do you like my avatar? Actually I don't look like this, I just liked it. You're Italian, right? What do you look like? Why don't you tell me something about yourself...

---

2 Good! I live in Philadelphia, Pennsylvania. It's about 45 miles southwest of New York City. I live in a suburb in Northwest Philly. We have a big house and a dog called Prince, a terrier. He's really cute! I'm a basketball fan. The Philadelphia Sixers are one of the best teams in the US. Where do you live? Tell me about your town...

---

3 That's cool! Are you still studying, then? Are you enjoying it? What's your major? I studied biology and I'm gonna be a zoologist. I totally love animals!! What would you like to do next? Any projects for next year?

---

4 Sounds good! Well, I just graduated from college and I'm taking a year off to travel around the world. I'm still trying to decide where to go. Have you travelled much? Anywhere you've been to you'd recommend? Maybe somewhere off the beaten track? Any ideas?

---

5 Fantastic! I'll go there for sure. I like traveling and meeting new people. Unfortunately I don't speak any foreign languages... not good when you want to meet the locals. I don't wanna end up speaking only with other American travelers. Do you like getting to know different people and cultures? Tell me about your friends from other countries...

---

6 Well, Europe is going to be my starting point, I got a flight from Philly to London. What is the one place in Italy I shouldn't miss? In your opinion, I mean. I can't go everywhere, a year is long, but not enough to see everything... ;-)

---

7 I was thinking I may go to Asia after Europe, from Moscow to India and then Nepal, Thailand and Sumatra. Cool, huh? I really wanna go someplace different, away from the American way of life... What is the best place you've ever been to? Where would you go if you could choose anywhere in the world? Why? Tell me...

---

8 What worries me most is the money. I don't think I have enough for a year and traveling is really expensive. I guess I could get a job in Europe, just to make money

before going East. Any advice? If I came to Italy, what could I do there? I don't speak Italian, though, what do you think?

---

9 Yeah, to tell you the truth, I don't mind working hard, in Philly I did some waitressing in a restaurant and in a bar. It can be tiring, but it's good money. And you? You got a job? What jobs are out there for you? With a major in languages getting a job should be easy...

---

10 You know what I'm really looking forward to? It's all the delicious food I'm gonna try. Here in Philly I eat out whenever I can. My favorite food is Thai, but I love Chinese and Japanese too. Do you like trying new dishes? What kind of food do you like most? I love Chow Mein, stir-fried noodles with chicken and vegetables. What's your favorite dish? Spaghetti meatballs? How often do you eat out?

---

11 All this talk about food is making me hungry already. I hear nightlife is great in Europe. I heard Spain is a crazy place, I guess Italy is the same. How do you spend your Saturday nights? What club do you go to in or around town? Do you stay out late? How do you get back home?

---

12 No night buses huh? Believe me when I say I don't like walking home on my own, but one can't always have a chaperone. I am afraid of the dark and silent side streets late at night, they give me the creeps. I always get the feeling I'm being followed. And you? What are you afraid of? Any special phobias? Like creepy critters, flying things, snakes?

---

13 They make me screech too. Well, I'm really enjoying chatting with you, you know. Thanks for that. I feel like we're friends now. Do you make friends over the Internet? Have you met anyone special through chats or Facebook? Are you on FB? Have you got many friends? 14 I have made many new friends through the Web. I really like chatting with them, like I'm enjoying chatting with you. Well, I better go now. Still have to book some flights and places to stay. Thanks for your help. We should chat again soon!! Maybe when I'm in Europe. ;-) bye

# Appendix II Word-sequences from LCC:

### Top twenty 2-word sequences:

| Rank | Raw Freq. | Norm. Freq. | 2-word seq. |
|---|---|---|---|
| 1 | 996 | 484.8 | i dont |
| 2 | 963 | 468.7 | and i |
| 3 | 935 | 455.1 | i have |
| 4 | 792 | 385.5 | i think |
| 5 | 778 | 378.7 | i like |
| 6 | 752 | 366 | but i |
| 7 | 722 | 351.4 | in the |
| 8 | 683 | 332.4 | a lot |
| 9 | 683 | 332.4 | i love |
| 10 | 642 | 312.5 | go to |
| 11 | 638 | 310.5 | like to |
| 12 | 625 | 304.2 | in a |
| 13 | 599 | 291.6 | you can |
| 14 | 550 | 267.7 | with my |
| 15 | 539 | 262.3 | if you |
| 16 | 517 | 251.6 | to go |
| 17 | 478 | 232.7 | i live |
| 18 | 473 | 230.2 | have a |
| 19 | 451 | 219.5 | live in |
| 20 | 451 | 219.5 | lot of |

**Top twenty 3-word sequences:**

| Rank | Raw Freq. | Norm. Freq. | 3-word seq. |
|---|---|---|---|
| 1 | 423 | 205.9 | a lot of |
| 2 | 362 | 176.2 | i live in |
| 3 | 267 | 130 | i think that |
| 4 | 259 | 126.1 | would like to |
| 5 | 250 | 121.7 | i dont like |
| 6 | 248 | 120.7 | i would like |
| 7 | 233 | 113.4 | i dont have |
| 8 | 223 | 108.5 | with my friends |
| 9 | 220 | 107.1 | to go to |
| 10 | 179 | 87.1 | i dont know |
| 11 | 166 | 80.8 | id like to |
| 12 | 163 | 79.3 | find a job |
| 13 | 154 | 75 | i have to |
| 14 | 153 | 74.5 | i have a |
| 15 | 153 | 74.5 | live in a |
| 16 | 139 | 67.7 | like to go |
| 17 | 134 | 65.2 | go to the |
| 18 | 134 | 65.2 | i really like |
| 19 | 124 | 60.4 | to find a |
| 20 | 121 | 58.9 | if you want |

**Top twenty 4-word sequences:**

| Rank | Raw Freq. | Norm. Freq. | 4-word seq. |
|---|---|---|---|
| 1 | 234 | 113.9 | i would like to |
| 2 | 138 | 67.2 | i live in a |
| 3 | 98 | 47.7 | to find a job |
| 4 | 84 | 40.9 | have a lot of |
| 5 | 80 | 38.9 | like to go to |
| 6 | 79 | 38.5 | i live in milan |
| 7 | 78 | 38 | out with my friends |
| 8 | 78 | 38 | there are a lot |
| 9 | 76 | 37 | are a lot of |
| 10 | 73 | 35.5 | if you want to |
| 11 | 68 | 33.1 | go out with my |
| 12 | 68 | 33.1 | i dont have any |
| 13 | 61 | 29.7 | all over the world |
| 14 | 59 | 28.7 | yes im still studying |
| 15 | 57 | 27.7 | to chat with you |
| 16 | 57 | 27.7 | would like to go |
| 17 | 56 | 27.3 | nice to meet you |
| 18 | 55 | 26.8 | i dont have a |
| 19 | 55 | 26.8 | the best place ive |
| 20 | 53 | 25.8 | i like very much |

**Top twenty 5-word sequences:**

| Rank | Raw Freq. | Normalised Freq. | 5-word seq. |
|---|---|---|---|
| 1 | 76 | 37 | there are a lot of |
| 2 | 48 | 23.4 | go out with my friends |
| 3 | 48 | 23.4 | i live in a small |
| 4 | 47 | 22.9 | i would like to go |
| 5 | 46 | 22.4 | i have a lot of |
| 6 | 40 | 19.5 | hi cherry my name is |
| 7 | 39 | 19 | would like to go to |
| 8 | 36 | 17.5 | in the north of italy |
| 9 | 34 | 16.5 | have a lot of friends |
| 10 | 34 | 16.5 | the best place ive ever |
| 11 | 33 | 16.1 | live in a small town |
| 12 | 31 | 15.1 | i go out with my |
| 13 | 30 | 14.6 | best place ive ever been |
| 14 | 29 | 14.1 | i would like to work |
| 15 | 29 | 14.1 | if you come to italy |
| 16 | 28 | 13.6 | cherry nice to meet you |
| 17 | 28 | 13.6 | make friends over the internet |
| 18 | 27 | 13.1 | hi cherry nice to meet |
| 19 | 26 | 12.7 | as i told you before |
| 20 | 25 | 12.2 | im years old and i |

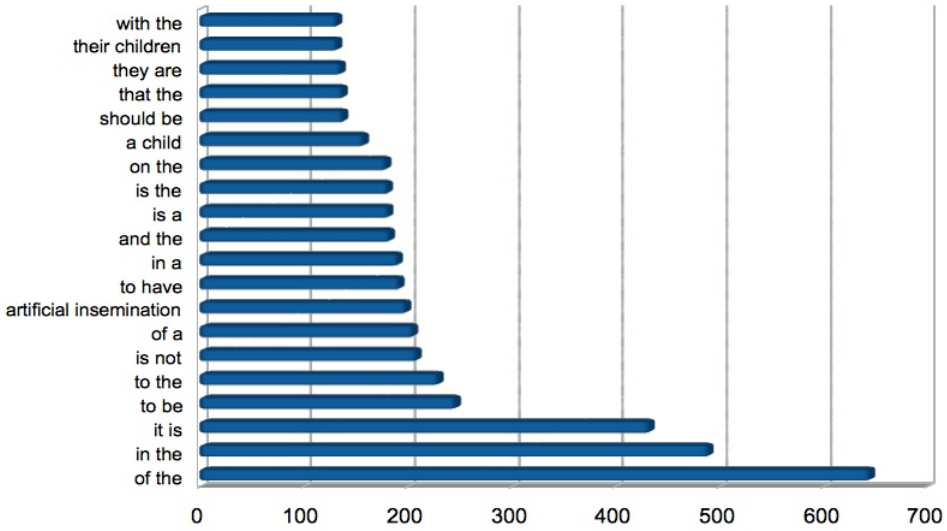
**6-word sequences:**

| Rank | Raw Freq. | Normalised Freq. | 6-word seq. |
|---|---|---|---|
| 1 | 32 | 15.6 | i live in a small town |
| 2 | 30 | 14.6 | i would like to go to |
| 3 | 30 | 14.6 | the best place ive ever been |
| 4 | 27 | 13.1 | i have a lot of friends |
| 5 | 26 | 12.7 | hi cherry nice to meet you |
| 6 | 22 | 10.7 | i go out with my friends |
| 7 | 20 | 9.7 | getting to know different people and |

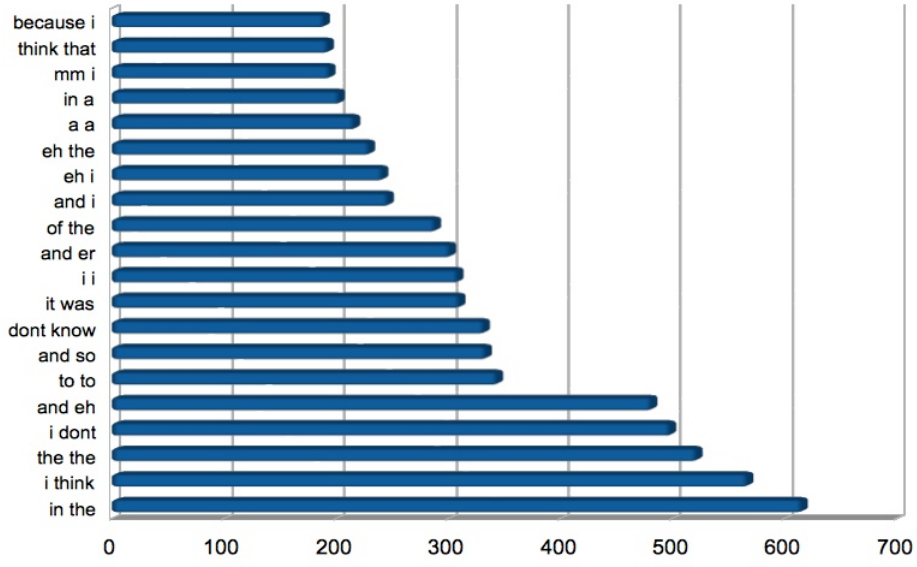
# Appendix III

## Sequences from ICLE_IT and LINDSEI_IT

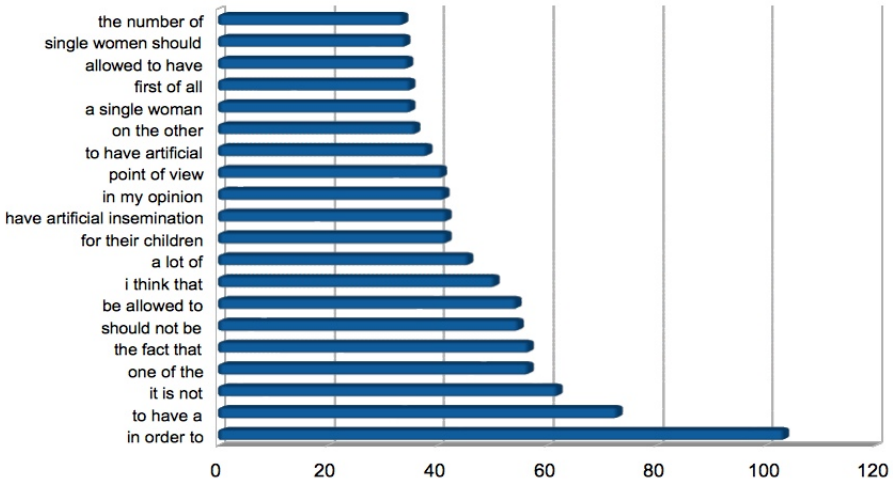
### 2-word Sequences: ICLE_IT



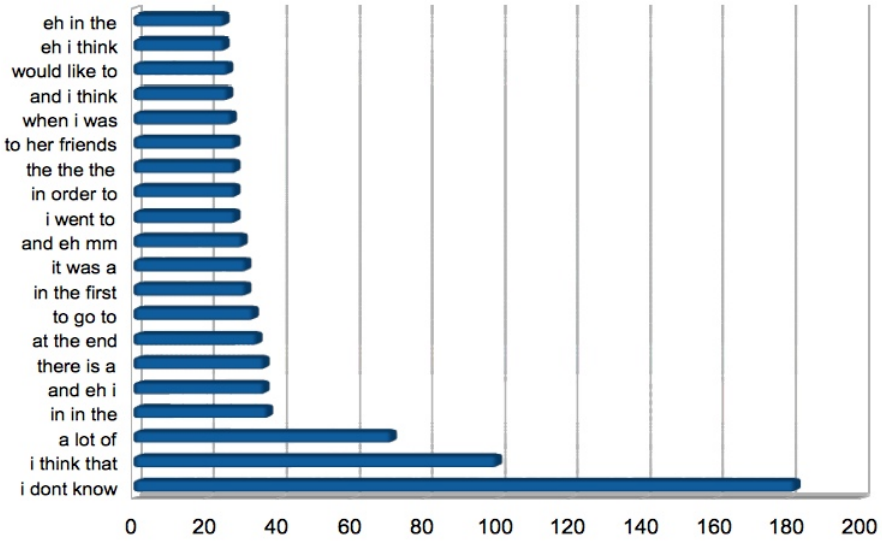

### 2-word Sequences: LINDSEI_IT



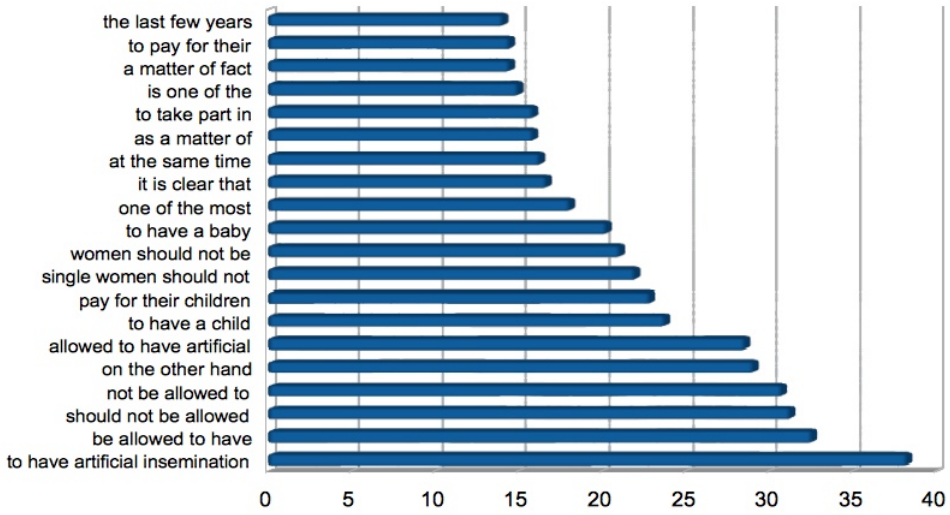
### 3-word Sequences: ICLE_IT



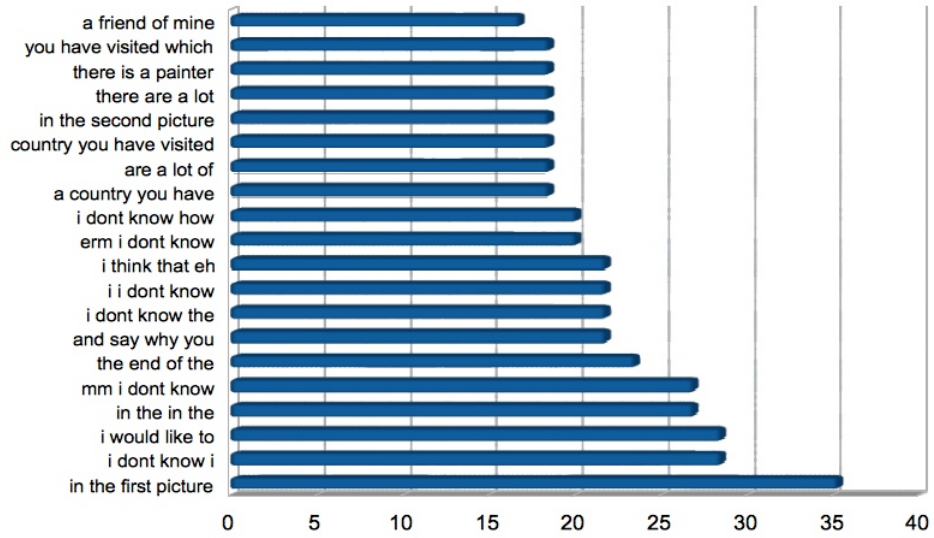

### 3-word Sequences: LINDSEI_IT



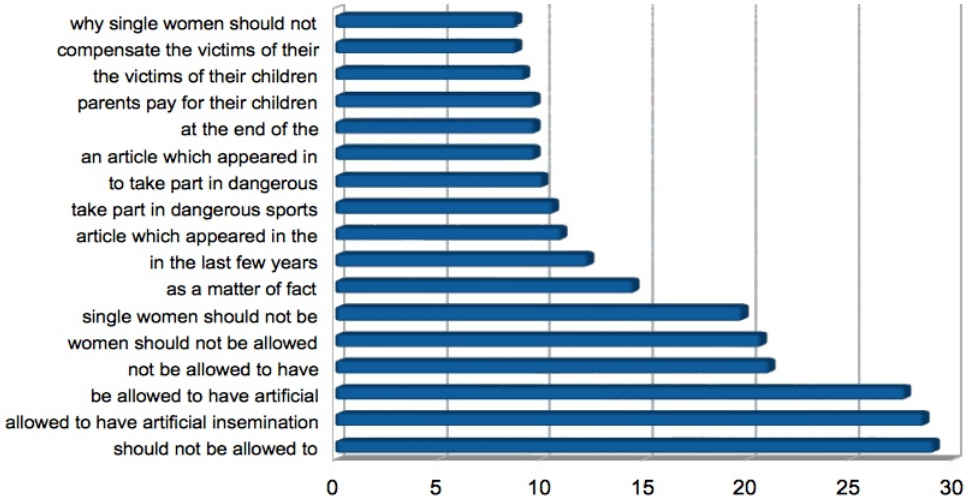
## 4-word Sequences: ICLE_IT




## 4-word Sequences: LINDSEI_IT

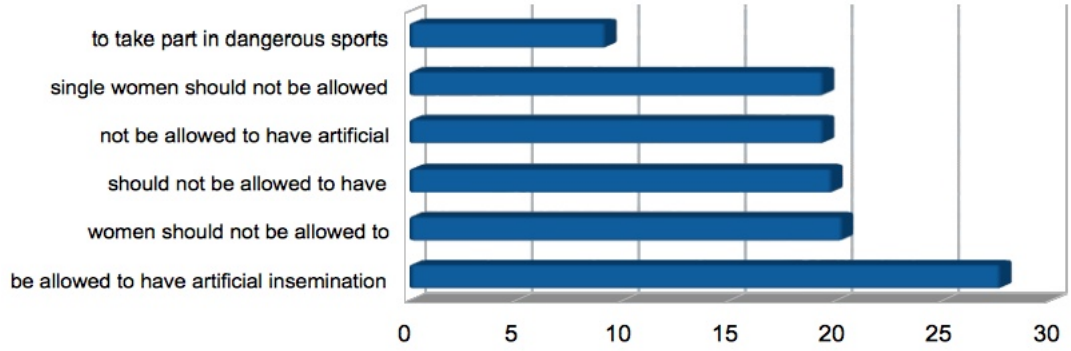


## 5-word Sequences: ICLE_IT




## 5-word Sequences: LINDSEI_IT

## 6-word Sequences: ICLE_IT




## 6-word Sequences: LINDSEI_IT